Independent Component Analysis Over Galois Fields

Arie Yeredor, *Senior Member, IEEE*

Abstract

We consider the framework of Independent Component Analysis (ICA) for the case where the independent sources and their linear mixtures all reside in a Galois field of prime order P. Similarities and differences from the classical ICA framework (over the Real field) are explored. We show that a necessary and sufficient identifiability condition is that none of the sources should have a Uniform distribution. We also show that pairwise independence of the mixtures implies their full mutual independence (namely a non-mixing condition) in the binary ($P = 2$) and ternary ($P = 3$) cases, but not necessarily in higher order ($P > 3$) cases. We propose two different iterative separation (or identification) algorithms: One is based on sequential identification of the smallest-entropy linear combinations of the mixtures, and is shown to be equivariant with respect to the mixing matrix; The other is based on sequential minimization of the pairwise mutual information measures. We provide some basic performance analysis for the binary ($P = 2$) case, supplemented by simulation results for higher orders, demonstrating advantages and disadvantages of the proposed separation approaches.

I. Introduction

Independent Component Analysis (ICA, see, e.g., [2], [3], [4] for some of the fundamental principles) addresses the recovery of unobserved, statistically independent source signals from their observed linear (and invertible) mixtures, without prior knowledge of the mixing matrix or of the sources' statistics. Classically, the ICA framework assumes that the sources and the mixing (hence, also the observations) are defined over the field of real-valued numbers $\mathbb{R}$, with some exceptions (e.g., [5]) that assume the field of complex-valued numbers $\mathbb{C}$. It might be interesting, though, at least from a theoretical point of view, to explore the applicability of ICA principles in other algebraic fields.

In this work we consider ICA over Galois Fields of prime order P, denoted $\mathbb{GF}(P)$, such that the sources and the mixing-matrix' elements can all take only a finite number of values, defined

The author is with the Department of Electrical Engineering - Systems, Tel-Aviv University, Tel-Aviv, Israel. Some part of this work was presented at ICA'07 [1].

by the set $\{0, 1, ..., P-1\}$ (or by some offset, isomorphic version thereof), and where addition and multiplication are applied modulu P, thereby returning values in the same set.

For example, in the field $\mathbb{GF}(2)$ of binary numbers $\{0, 1\}$, addition is obviously equivalent to the "Exclusive Or" (XOR) operation, denoted $z = x \oplus y$ (where z equals 1 if $x \neq y$ and equals 0 otherwise). Multiplication (either by 0 or by 1) is defined in the "usual" way in this case.

In the field $\mathbb{GF}(3)$ of ternary numbers $\{0, 1, 2\}$, where addition and multiplication are defined modulu 3 (similarly denoted $z = x \oplus y$), it is sometimes more convenient to consider the offset group $\{0, 1, -1\}$. In this group, multiplication can still be defined in the "usual" way, since ordinary multiplication of any two numbers in this group returns a number in the group. Obviously, the two sets $\{0, 1, 2\}$ and $\{0, 1, -1\}$ are isomorphic in $\mathbb{GF}(3)$, and will be used interchangeably in the sequel.

A fundamental difference, at least in the context of ICA, between random variables over $\mathbb{R}$ and over $\mathbb{GF}(P)$ is the following: Let u and v be two statistically independent, non-degenerate (namely, non-deterministic) random variables, and consider the random variable w, given by any non-trivial linear combination of u and v. In $\mathbb{R}$, v and w cannot be statistically independent (they are obviously correlated), no matter how u and v are distributed. However, as we shall show in Section III, in $\mathbb{GF}(P)$ v and w may indeed be statistically independent, and this happens if and only if the distribution of u is uniform (taking each of the P values with equal probabilities).

In a sense, this property tags the uniform distribution as the "problematic" distribution in ICA over $\mathbb{GF}(P)$, reminiscent of the role taken by the Gaussian distribution in ICA over $\mathbb{R}$. Note that these two distributions share additional related properties in their respective fields: They are both (under mild regularity conditions) limit-distributions of an infinite sum of independent random variables; and they are both "maximum entropy" distributions (subject to a variance constraint for the Gaussian distribution in $\mathbb{R}$). So, loosely stated, in the same way that a linear combination of independent random variables over $\mathbb{R}$ tends to be "more Gaussian", a linear combination of independent random variables over $\mathbb{GF}(P)$ tends to be "more uniform".

Nevertheless, there still remain some essential differences between the roles of these distributions in the respective contexts. For example, in $\mathbb{GF}(P)$, if (at least) one of random variables in the linear combination of independent variables is uniform, the resulting distribution would be exactly uniform as well, no matter how the other random variables are distributed. Evidently, this property does not hold for Gaussian distributions over $\mathbb{R}$.

Therefore, as we shall show, these properties lead to an identifiability condition for ICA over $\mathbb{GF}(P)$, which is reminiscent of, but certainly not equivalent to, a well-known identifiability condition over $\mathbb{R}$. More specifically, the identifiability condition for ICA over $\mathbb{R}$ requires that *not more than one* of the sources be Gaussian. Our identifiability condition for ICA over $\mathbb{GF}(P)$ requires that *none* of the sources be uniform. The key to this identifiability condition is the property that the entropy of any linear combination of statistically independent random variables over $\mathbb{GF}(P)$ is larger than the entropy of the largest-entropy component, as long as this component is not uniform. Therefore, if none of the sources is uniform, then, at least conceptually, a possible separation approach is to look for the (inverse) linear transformation, which minimizes the empirical marginal entropies of the resulting linear combinations. However, since an exhaustive search for this transformation would often be prohibitively computationally expensive, we shall propose an alternative, computationally cheaper method for entropy-based identification.

Another possible, somewhat different separation approach is the following. One of the key observations in ICA over $\mathbb{R}$ is that, under the identifiability condition and due to the Darmois-Skitovitch theorem (e.g., [6], p.218), pairwise-independence of the mixtures implies their full mutual independence, which in turn implies a non-mixing condition (namely, separation). Interestingly, we shall show that our general identifiability condition is necessary and sufficient to guarantee a similar property for ICA over $\mathbb{GF}(2)$ and $\mathbb{GF}(3)$, but is generally insufficient for this property to hold in $\mathbb{GF}(P)$ for $P > 3$. Thus, another possible identification approach (in $\mathbb{GF}(2)$ and in $\mathbb{GF}(3)$ only) is to look for an invertible linear transformation of the observations, which makes the resulting signals "as empirically pairwise-independent as possible" - a property which is easier to quantify and measure than full independence (being quadratic, rather than exponential, in the number of sources K). Again - since an exhaustive search is often not feasible, we shall propose a different, sequential method for this approach.

A common assumption in the design and analysis of classical ICA methods over $\mathbb{R}$, is that each of the sources has an independent, identically distributed (iid) time-structure. Our discussion in this paper would be similarly restricted along the same line. We note, however, that in equivalence to methods which exploit possibly different temporal structures (e.g., spectral diversity [7], non-stationarity [8], etc.) over $\mathbb{R}$, similar extensions of our results would be possible in similar cases over $\mathbb{GF}(P)$. However, we defer the exploration of such cases to future work.

The paper is structured as follows. In the next section we review some fundamental properties of random variables and random vectors in $\mathbb{GF}(P)$, which will be useful in subsequent derivations. In Section III we outline the problem formulation and present our general identifiability condition. In Section IV we explore the relation between pairwise independence and full independence, showing that in an invertible linear mixture, the former implies the latter in $\mathbb{GF}(2)$ and in $\mathbb{GF}(3)$, but not necessarily in Galois fields of higher orders. We then proceed to propose two different separation algorithm in Section V. A rudimentary performance analysis for the simple binary case ($P = 2$) is provided in Section VI, supplemented with supporting simulation results which extend to larger-scale scenarios. Our work is summarized with concluding remarks in Section VII.

We shall denote addition, subtraction and multiplication over $\mathbb{GF}(P)$ (namely, modulu P) by $\oplus$, $\ominus$

and $\otimes$, respectively, with multiplication preceding addition and subtraction in the order of operations. Vector multiplication will be denoted by $\circ$, such that if $\boldsymbol{a} = [a_1 \ \cdots \ a_K]^T$ and $\boldsymbol{x} = [x_1 \ \cdots \ x_K]^T$, then

$$\boldsymbol{a}^T \circ \boldsymbol{x} = a_1 \otimes x_1 \oplus a_2 \otimes x_2 \oplus \cdots \oplus a_K \otimes x_K. \tag{1}$$

Similarly, if $\boldsymbol{A}$ is an $L \times K$ matrix in $\mathbb{GF}(P)$, its product with $\boldsymbol{x}$ is denoted $\boldsymbol{A} \circ \boldsymbol{x}$, an $L \times 1$ vector whose elements are the products of the respective rows of $\boldsymbol{A}$ with $\boldsymbol{x}$.

II. Characterization of random variables and random vectors in $\mathbb{GF}(P)$

We begin by briefly outlining some of the basic essential properties and definitions of our notations for random variables and random vectors in $\mathbb{GF}(P)$, which we shall use in the sequel.

A random variable u in $\mathbb{GF}(P)$ is characterized by a discrete probability distribution, fully described by a vector $\boldsymbol{p}_u = [p_u(0) \ p_u(1) \ \cdots \ p_u(P-1)]^T \in \mathbb{R}^P$, whose elements $p_u(m)$ are $\Pr\{u = m\}$, the probabilities of u taking the values $m \in \{0, \ldots, P-1\}$. Evidently, all the elements of $\boldsymbol{p}_u$ are non-negative and their sum equals 1. We shall refer to $\boldsymbol{p}_u$ as the *probability vector* of u. The *entropy* of u is given by

$$H(u) = -\sum_{m=0}^{P-1} p_u(m) \log p_u(m). \tag{2}$$

By maximizing with respect to $\boldsymbol{p}_u$, it is easy to show that among all random variables in $\mathbb{GF}(P)$, the uniform random variable (taking all values in $\mathbb{GF}(P)$ with equal probability $\frac{1}{P}$) has the largest entropy, given by $\log P$. Note that it is convenient to use a base-P logarithm $\log_P$ (rather than the more commonly-used $\log_2$) in this context, such that the entropies of all (scalar) random variables in $\mathbb{GF}(P)$ are confined to $[0, 1]$. Note, in addition, that since multiplication by a constant over $\mathbb{GF}(P)$ is bijective, the entropy of a random variable in $\mathbb{GF}(P)$ is invariant under such multiplication (which merely re-arranges the terms in the sum (2)).

The *characteristic vector* of u is denoted $\tilde{\boldsymbol{p}}_u = [\tilde{p}_u(0) \ \tilde{p}_u(1) \ \cdots \ \tilde{p}_u(P-1)]^T \in \mathbb{C}^P$, and its elements are given by the discrete Fourier transform (DFT) of the elements of $\boldsymbol{p}$:

$$\tilde{p}_u(n) = E[W_P^{nu}] = \sum_{m=0}^{P-1} p_u(m) W_P^{mn} \quad n = 0, \ldots, P-1, \tag{3}$$

where the “twiddle factor” W_P is defined as $W_P = e^{-j2\pi/P}$ (note that the modulu-P operation is inherently present in the exponential part, so W_P^{mn} is equivalent to $W_P^{m \otimes n}$). Like the probability vector $\boldsymbol{p}_u$, the characteristic vector $\tilde{\boldsymbol{p}}_u$ provides full statistical characterization of the random variable u, since $\boldsymbol{p}_u$ can be directly obtained from $\tilde{\boldsymbol{p}}_u$ using the inverse DFT.

The following basic properties of $\tilde{\boldsymbol{p}}_u$ can be easily obtained:

P1) $\tilde{p}_u(0) = 1$;

P2) Since $\boldsymbol{p}_u$ is real-valued, $\tilde{p}_u(n) = \tilde{p}_u^*(P-n)$ (where the superscript * denotes the complex-conjugate);

P3) u is uniform (namely, $p_u(m) = \frac{1}{P}\ \forall m$) $\Leftrightarrow \tilde{p}_u(n) = 0\ \forall n \neq 0$;

P4) u is degenerate (namely, $p_u(M) = 1$ for some M) $\Leftrightarrow \tilde{p}_u(n) = W_P^{nM}\ \forall n$;

P5) $|\tilde{p}_u(n)| \leq 1\ \forall n$, where for $n \neq 0$ equality holds if and only if (iff) u is degenerate.

Note that in the particular cases of $\mathbb{GF}(2)$ and $\mathbb{GF}(3)$ we have the following simplifications:

- In $\mathbb{GF}(2)$, the only free parameter in $\tilde{\boldsymbol{p}}_u \in \mathbb{R}^2$ is $\tilde{p}_u(1)$, to which we shall refer as

$$\theta_u \overset{\triangle}{=} \tilde{p}_u(1) = p_u(0) - p_u(1) = 1 - 2p_u(1). \tag{4}$$

  Thus $\tilde{\boldsymbol{p}}_u = [1\ \theta_u]^T$;

- In $\mathbb{GF}(3)$, there is also a single (yet complex-valued) free parameter in $\tilde{\boldsymbol{p}}_u \in \mathbb{C}^3$, to which we shall refer as

$$\xi_u \overset{\triangle}{=} \tilde{p}_u(1) = p_u(0) + p_u(1)W_3^{-1} + p_u(2)W_3^{-2} = 1 - \tfrac{3}{2}(p_u(1)+p_u(2)) + j\tfrac{\sqrt{3}}{2}(p_u(2)-p_u(1)). \tag{5}$$

  Thus $\tilde{\boldsymbol{p}}_u = [1\ \xi_u\ \xi_u^*]^T$.

Note also that $\theta_u = E[W_2^u] = E[(-1)^u]$ and $\xi_u = E[W_3^u]$.

For two random variables u and v in $\mathbb{GF}(P)$, the joint statistics are completely described by the *joint probabilities matrix* $\boldsymbol{P}_{u,v} \in \mathbb{R}^{P\times P}$, whose elements are $P_{u,v}(m,n) = \Pr\{u = m, v = n\}$, $m,n \in \{0,\ldots,P-1\}$. The joint entropy of u and v is given by

$$H(u,v) = -\sum_{m,n=0}^{P-1} P_{u,v}(m,n)\log P_{u,v}(m,n). \tag{6}$$

The random variables u and v are said to be *statistically independent* iff $\boldsymbol{P}_{u,v} = \boldsymbol{p}_u\boldsymbol{p}_v^T$. By Jensen's inequality, $H(u,v)$ satisfies $H(u,v) \leq H(u) + H(v)$, with equality iff u and v are statistically independent. The *mutual information* between u and v is the difference $I(u,v) \overset{\triangle}{=} H(u)+H(v)-H(u,v)$, which is also the (non-negative) Kullback-Leibler divergence between their joint distribution and the product of their marginal distributions. The smaller their mutual information, the “more statistically independent” u and v are; $I(u,v)$ vanishes if and only if u and v are statistically independent.

The conditional distribution of u given v is given by $\boldsymbol{P}_{u|v} \in \mathbb{R}^{P\times P}$ with elements $P_{u|v}(m,n) = P_{u,v}(m,n)/p_v(n) = \Pr\{u = m|v = n\}$, $m,n = 0,\ldots,P-1$. The conditional entropy is defined as

$$H(u|v) = -\sum_{n=0}^{P-1} p_v(n)\sum_{m=1}^{P-1} P_{u|v}(m,n)\log P_{u|v}(m,n), \tag{7}$$

which can be easily shown to satisfy $H(u|v) = H(u,v) - H(v)$.

The joint *characteristic matrix* of u and v, denoted $\widetilde{\boldsymbol{P}}_{u,v} \in \mathbb{C}^{P\times P}$, is given by the two-dimensional DFT (2DFT) of $\boldsymbol{P}_{u,v}$,

$$\widetilde{P}_{u,v}(m,n) = E[W_P^{mu+nv}] = \sum_{k,\ell=0}^{P-1} P_{u,v}(k,\ell)W_P^{mk+n\ell}, \tag{8}$$

and provides an alternative full statistical characterization of u and v. In particular, it is straightforward to show that $\widetilde{\boldsymbol{P}}_{u,v}$ satisfies $\widetilde{\boldsymbol{P}}_{u,v} = \tilde{\boldsymbol{p}}_u \tilde{\boldsymbol{p}}_v^T$ iff u and v are statistically independent.

For a $K \times 1$ random vector $\boldsymbol{u}$ whose elements $u_1, \ldots, u_K$ are random variables in $\mathbb{GF}(P)$, the joint statistics are fully characterized by the K-way *probabilities tensor* $\boldsymbol{\mathcal{P}}_{\boldsymbol{u}} \in \mathbb{R}^{P^{(\times K)}}$, whose elements are the probabilities $\mathcal{P}_{\boldsymbol{u}}(m_1, \ldots, m_K) = \Pr\{u_1 = m_1, \ldots, u_K = m_K\}$, $m_1, \ldots, m_K \in \{0, \ldots, P-1\}$. Using vector-index notations, where $\boldsymbol{m} = [m_1, \cdots, m_K]^T$, we may also express this relation more compactly as $\mathcal{P}_{\boldsymbol{u}}(\boldsymbol{m}) = \Pr\{\boldsymbol{u} = \boldsymbol{m}\}$. The *characteristic tensor* $\widetilde{\boldsymbol{\mathcal{P}}}_{\boldsymbol{u}} \in \mathbb{C}^{P^{(\times K)}}$ is given by the K-dimesional DFT of $\boldsymbol{\mathcal{P}}_{\boldsymbol{u}}$, which, using a similar index-vector notation, is given by

$$\widetilde{\mathcal{P}}_{\boldsymbol{u}}(\boldsymbol{n}) = E[W_P^{\boldsymbol{n}^T \boldsymbol{u}}] = \sum_{\boldsymbol{m}} \mathcal{P}_{\boldsymbol{u}}(\boldsymbol{m}) W_P^{\boldsymbol{n}^T \boldsymbol{m}}. \tag{9}$$

where the summation extends over all possible P^K indices combinations in $\boldsymbol{m}$.

III. Problem Formulation and Indentifiability

We are now ready to formulate the mixture model over $\mathbb{GF}(P)$. Assume that there are K statistically independent random source signals denoted $\boldsymbol{s}[t] = [s_1[t] \;\; s_2[t] \;\; \cdots \;\; s_K[t]]^T$, each with an iid time-structure, such that at each time-instant t, $s_k[t]$ is an independent realizations of a random variable in $\mathbb{GF}(P)$, characterized by the (unknown) distribution vector $\boldsymbol{p}_k$.

Let these sources be mixed (over $\mathbb{GF}(P)$) by an unknown, square ($K \times K$) mixing matrix $\boldsymbol{A}$ (with elements in $\mathbb{GF}(P)$),

$$\boldsymbol{x}[t] = \boldsymbol{A} \circ \boldsymbol{s}[t]. \tag{10}$$

We further assume that $\boldsymbol{A}$ is invertible over the field, namely that it has a unique inverse over $\mathbb{GF}(P)$, denoted $\boldsymbol{B} \triangleq \boldsymbol{A}^{-1}$, satisfying $\boldsymbol{B} \circ \boldsymbol{A} = \boldsymbol{A} \circ \boldsymbol{B} = \boldsymbol{I}$, where $\boldsymbol{I}$ denotes the $K \times K$ identity matrix. Like in "classical" linear algebra (over $\mathbb{R}$), $\boldsymbol{A}$ is non-singular (invertible) iff its determinant[1] is non-zero. Equivalently, $\boldsymbol{A}$ is singular iff there exists (in $\mathbb{GF}(P)$) a nonzero vector $\boldsymbol{u}$, such that $\boldsymbol{A} \circ \boldsymbol{u} = \boldsymbol{0}$ (an all-zeros vector).

We are interested in the identifiability, possibly up to some tolerable ambiguities, of $\boldsymbol{A}$ (or, equivalently, of its inverse $\boldsymbol{B}$) from the set of observations $\boldsymbol{x}[t]$, $t = 1, 2, ...T$ under asymptotic conditions, namely as $T \to \infty$. Due to the assumption of iid samples for each source (implying ergodicity), the joint statistics of the observations can be fully and consistently estimated from the available data. Therefore, the assumption of asymptotic conditions implies full and exact knowledge of the joint probability distribution tensor $\boldsymbol{\mathcal{P}}_{\boldsymbol{x}}$ of the observation vector $\boldsymbol{x}$ (we dropped the time-index

[1] The determinant over $\mathbb{GF}(P)$ can be calculated in a similar way to calculating the determinant over $\mathbb{R}$, using the field's addition/subtraction and multiplication operations.

t here, due to the stationarity). The remaining question is, therefore - whether, and if so, under what conditions, $\boldsymbol{A}$ can be identified (up to tolerable ambiguities) from exact, full knowledge of $\boldsymbol{\mathcal{P}_x}$.

To answer this question, we first explore some basic statistical properties of linear combinations of random variables over $\mathbb{GF}(P)$. The characteristic vectors are particularly useful for this analysis. Let u and v denote two statistically independent random variables in $\mathbb{GF}(P)$ with probability vectors $\boldsymbol{p}_u$ and $\boldsymbol{p}_v$ and characteristic vectors $\tilde{\boldsymbol{p}}_u$ and $\tilde{\boldsymbol{p}}_v$, respectively. If $w = u \oplus v$, then the probability vector $\boldsymbol{p}_w$ of w is given by the cyclic convolution between $\boldsymbol{p}_u$ and $\boldsymbol{p}_v$, and the characteristic vector $\tilde{\boldsymbol{p}}_w$ is therefore given by the element-wise product of $\tilde{\boldsymbol{p}}_u$ and $\tilde{\boldsymbol{p}}_v$:

$$p_w(n) = \sum_{m=0}^{P-1} \Pr\{u = m, v = n \ominus m\} = \sum_{m=0}^{P-1} p_u(m) p_v(n \ominus m) \quad \Leftrightarrow \quad \tilde{p}_w(n) = \tilde{p}_u(n)\tilde{p}_v(n) \quad \forall n. \tag{11}$$

Two intuitively appealing (nearly trivial) properties follow from this relation. First, combined with Property P4 (in Section II), this relation implies that the sum (over $\mathbb{GF}(P)$) of two independent random variables is a degenerate random variable iff both are degenerate. Likewise, combined with Property P3, this relation implies that the sum is uniform if at least one of the variables is uniform. The converse, however, is perhaps somewhat less trivial, since it involves a distinction between $\mathbb{GF}(2)$ and $\mathbb{GF}(3)$ on one hand, and $\mathbb{GF}(P)$ with $P > 3$ on the other hand, as suggested by the following lemma:

Lemma 1: Let u and v be two statistically independent random variables in $\mathbb{GF}(P)$, and let $w \triangleq u \oplus v$. If both u and v are non-uniform, then:

1) If $P = 2$ or $P = 3$, w is also non-uniform;
2) If $P > 3$, w may or may not be uniform.

Proof: By Property P3, w would be uniform iff for each $n \neq 0$, either $\tilde{p}_u(n) = 0$ or $\tilde{p}_v(n) = 0$ (or both). In $\mathbb{GF}(2)$ this can only happen if either $\theta_u = 0$ or $\theta_v = 0$ (or both), which implies that at least one of the two variables is uniform. Likewise, in $\mathbb{GF}(3)$ this can only happen if either $\xi_u = 0$ or $\xi_v = 0$ (or both), leading to a similar conclusion.

However, for $P > 3$ there are sufficiently many degrees of freedom in the characteristic vectors of u and v to allow both non-zero and zero elements in both $\tilde{\boldsymbol{p}}_u$ and $\tilde{\boldsymbol{p}}_v$, as long as at each $n \neq 0$ either one is zero. For example, consider $P = 5$ with $\tilde{\boldsymbol{p}}_u = [1\ 0\ 0.3\ 0.3\ 0]^T$ and $\tilde{\boldsymbol{p}}_v = [1\ 0.4\ 0\ 0\ 0.4]^T$. This corresponds to $\boldsymbol{p}_u \approx [0.32\ 0.10\ 0.24\ 0.24\ 0.10]^T$ and $\boldsymbol{p}_v \approx [0.36\ 0.25\ 0.07\ 0.07\ 0.25]^T$, which are clearly non-uniform. However, if these u and v are independent, their sum (over $\mathbb{GF}(5)$) is a uniform random variable. ■

Note, in addition, that since multiplication by a constant in $\mathbb{GF}(P)$ is bijective, uniform or degenerate random variables cannot become non-uniform or non-degenerate (nor vice-versa) by multiplication with a constant. Consequently, the above conclusions and Lemma 1 hold not only for the *sum* of two random variables, but also for any *linear combination* (over $\mathbb{GF}(P)$) thereof.

We now add the following Lemma:

Lemma 2: Let u and v be two statistically independent, non-degenerate random variables in $\mathbb{GF}(P)$, and let $w \overset{\triangle}{=} u \oplus v$. Then v and w are statistically independent iff u is uniform.

Proof: The joint probability distribution of v and w is given by

$$P_{v,w}(m,n) = \Pr\{v=m, w=n\} = \Pr\{v=m, u=n\ominus m\} = p_v(m)p_u(n\ominus m). \tag{12}$$

Now, w and v are independent iff this probability equals $p_v(m)p_w(n)$ for all m,n, namely iff $p_u(n\ominus m) = p_w(n)$ for all n and for all m with which $p_v(m) \neq 0$. Since v is non-degenerate, there are at least two such values of m. Denoting these values as m_1 and m_2, this condition translates into

$$p_u(n\ominus m_1) = p_u(n\ominus m_2) = p_w(n) \quad \forall n. \tag{13}$$

We therefore also have $p_u(n) = p_u(n\oplus m_1 \ominus m_2)\ \forall n$, which can be recursively generalized into

$$p_u(n) = p_u(n\oplus k\otimes(m_1\ominus m_2)) \quad \forall n,k \in \mathbb{GF}(P). \tag{14}$$

Since P is prime, each element in $\mathbb{GF}(P)$ can be represented (given n, m_1 and m_2) as $n\oplus k\otimes(m_1\ominus m_2)$ with some k, therefore this condition is satisfied iff $p_u(n)$ is constant, namely iff u is uniform. ■

To establish our identifiability condition we need one additional lemma, which characterizes the entropy of a linear combination of random variables in $\mathbb{GF}(P)$.

Lemma 3: Let u and v be two statistically independent, non-degenerate random variables in $\mathbb{GF}(P)$, and let $w \overset{\triangle}{=} u \oplus v$. Then $H(w) \geq H(u)$, where equality holds iff u is uniform.

Proof: As already mentioned in Section II, $H(w,v) \leq H(w)+H(v)$, with equality iff w and v are statistically independent. In addition, $H(w|v) = H(w,v) - H(v)$. Therefore, $H(w|v) \leq H(w)$, with equality iff w and v are statistically independent. Next, from (12) we have $\boldsymbol{P}_{w|v}(m,n) = p_u(n\ominus m)$, and therefore, as could be intuitively expected,

$$H(w|v) = \sum_{m=0}^{P-1} p_v(m) \sum_{n=0}^{P-1} p_u(n\ominus m)\log p_u(n\ominus m) = \sum_{m=0}^{P-1} p_v(m)H(u) = H(u), \tag{15}$$

and we therefore conclude that $H(u) \leq H(w)$, with equality iff w and v are statistically independent. Now, according to Lemma 2, w and v are statistically independent iff u is uniform, which completes the proof. ■

Obviously, a similar result (namely $H(w) \geq H(v)$) can be obtained by switching roles between u and v in the proof. Note an essential difference from a similar result over $\mathbb{R}$: In $\mathbb{R}$ the entropy (or differential entropy) of a sum of two independent, non-degenerate random variables is *always* strictly larger than their individual entropies, no matter how they are distributed. In $\mathbb{GF}(P)$, however, equality is attained if one of the variables is uniform. In fact, this equality is inevitable, simply because the entropy of any random variable in $\mathbb{GF}(P)$ is upper-bounded by the uniform variable's entropy (of $\log P$).

We are now ready to state our identifiability condition:

Theorem 1: Let $\boldsymbol{s}$ be a $K \times 1$ random vector whose elements are statistically-independent, non-degenerate random variables in $\mathbb{GF}(P)$. Let $\boldsymbol{A}$ be a $K \times K$ non-singular matrix in $\mathbb{GF}(P)$, and let the random vector $\boldsymbol{x}$ be defined as $\boldsymbol{x} = \boldsymbol{A} \circ \boldsymbol{s}$. Assume that the probability distribution of $\boldsymbol{x}$ is fully known (specified by the probabilities tensor $\boldsymbol{\mathcal{P}}_{\boldsymbol{x}}$). Then $\boldsymbol{A}$ can be identified, up to possible permutation and scaling of its columns, from $\boldsymbol{\mathcal{P}}_{\boldsymbol{x}}$ alone, iff *none of the elements of $\boldsymbol{s}$ is a uniform random variable.*

Proof: The necessity of this condition is obvious by Lemma 2. Even in the simplest 2×2 case, if one of the sources, say s_1, is uniform, then by Lemma 2 any linear combination of s_1 with the other source s_2 is still statistically independent of s_2. Therefore, if the mixed signals are $x_1 = s_1 \oplus s_2$ and $x_2 = s_2$, then x_1 and x_2 are statistically independent - so this situation is indistinguishable from a non-mixing observation of two independent sources with the same marginal distributions as x_1 and x_2 (which are also the marginal distributions of s_1 and s_2 (resp.) in this case).

To observe the sufficiency of the condition, note first that since $\boldsymbol{A}$ is invertible over $\mathbb{GF}(P)$, any invertible linear mixture of the original sources $\boldsymbol{s}$ can be obtained by applying some invertible linear mixing to the observations $\boldsymbol{x}$. Therefore, by applying all (finite number of) invertible linear transformations to $\boldsymbol{x}$, one can implicitly obtain all the invertible linear transformations of $\boldsymbol{s}$. Indeed, let $\hat{\boldsymbol{B}}$ denote an arbitrary invertible matrix in $\mathbb{GF}(P)$, and denote

$$\boldsymbol{y} \overset{\triangle}{=} \hat{\boldsymbol{B}} \circ \boldsymbol{x} = (\hat{\boldsymbol{B}} \circ \boldsymbol{A}) \circ \boldsymbol{s} \tag{16}$$

Since both $\hat{\boldsymbol{B}}$ and $\boldsymbol{A}$ are non-singular, so is $\hat{\boldsymbol{B}} \circ \boldsymbol{A}$, which therefore:

1) Has at least one non-zero element in each row; and
2) Has at least one non-zero element in each column, which means that each element of $\boldsymbol{s}$ is a component of (namely, participates with nonzero weight in) at least one element of $\boldsymbol{y}$.

Now define the respective sums of (marginal) entropies, $H_{mar}(\boldsymbol{y}) \overset{\triangle}{=} \sum_{k=1}^{K} H(y_k)$ and $H_{mar}(\boldsymbol{s}) \overset{\triangle}{=} \sum_{k=1}^{K} H(s_k)$. Consequently, by Lemma 3, $H_{mar}(\boldsymbol{y})$ cannot be made smaller than $H_{mar}(\boldsymbol{s})$. Moreover, if none of the elements of $\boldsymbol{s}$ is uniform, then

$$H_{mar}(\boldsymbol{y}) = H_{mar}(\boldsymbol{s}) \quad \Leftrightarrow \quad \hat{\boldsymbol{B}} \circ \boldsymbol{A} = \boldsymbol{\Pi} \circ \boldsymbol{\Lambda}, \tag{17}$$

where $\boldsymbol{\Pi}$ denotes a $K \times K$ permutation matrix and $\boldsymbol{\Lambda}$ denotes a $K \times K$ diagonal, nonsingular matrix in $\mathbb{GF}(P)$. Any other form of $\hat{\boldsymbol{B}} \circ \boldsymbol{A}$ would imply that at least one of the elements of $\boldsymbol{y}$ is a linear combination of at least two elements of $\boldsymbol{s}$, and as such has higher entropy than both, and since at least one of these two elements is also present in at least one other element of $\boldsymbol{y}$, $H_{mar}(\boldsymbol{y})$ must be larger than $H_{mar}(\boldsymbol{s})$.

It is therefore possible, at least conceptually, to apply each $K \times K$ nonsingular matrix $\hat{\boldsymbol{B}}$ in $\mathbb{GF}(P)$ to $\boldsymbol{x}$, and select one of the minimizers of $H_{mar}(\boldsymbol{y})$. The inverse of this minimizer is guaranteed to be equivalent to $\boldsymbol{A}$ up to permutation and scaling,

$$\hat{\boldsymbol{B}} \circ \boldsymbol{A} = \boldsymbol{\Pi} \circ \boldsymbol{\Lambda} \quad \Leftrightarrow \quad \hat{\boldsymbol{B}}^{-1} = \boldsymbol{A} \circ \boldsymbol{\Lambda}^{-1} \circ \boldsymbol{\Pi}^{T} \tag{18}$$

(where all the inverses are obviously taken over $\mathbb{GF}(P)$). ■

Note that in $\mathbb{GF}(2)$ the scaling ambiguity is meaningless, because the only possible scalar multiplication is by 1, therefore only the permutation ambiguity remains. In $\mathbb{GF}(3)$ the possible scaling ambiguity entails multiplication by either 1 or 2, or, if the "offset group" $\{0, 1, -1\}$ is used, this ambiguity merely translates into a sign-ambiguity.

Although the number of $K \times K$ nonsingular matrices in $\mathbb{GF}(P)$ is finite, this number is of the order of $P^{(K^2)}$, which clearly becomes prohibitively large even with relatively small values of P and K. Therefore, our identifiability proof, which is based on an exhaustive search, can hardly be translated into a practical separation scheme. Nevertheless, in Section V below we shall propose and discuss two practical separation approaches, which require a significantly reduced computational effort. First, however, we need to address one more theoretical aspect of our model - which is: whether (and if so under what conditions) pairwise independence of linear mixtures implies their full mutual independence.

IV. Pairwise independence implying full independence

One of the basic, key concepts in ICA over $\mathbb{R}$ is the Darmois-Skitovich Theorem (e.g., [6] p.218), which is used, either explicitly or implicitly, in many ICA methods ([2]). This theorem states that if two linear combinations (over $\mathbb{R}$) of statistically independent random variables are statistically independent, then all the random variables which participate (with non-zero coefficients) in both combinations must be Gaussian. Consequently (see, e.g., [2]), under the classical identifiability condition (for ICA over $\mathbb{R}$) of not more than one Gaussian source, pairwise statistical independence of linear mixtures of the sources always implies their full mutual statistical independence (namely, a non-mixing condition).

As we shall show in this section, this property does not carry over to our $\mathbb{GF}(P)$ scenario by mere substitution of the Gaussian distribution with the uniform. As it turns out, under our identifiability condition (for ICA over $\mathbb{GF}(P)$) of no uniform sources, pairwise independence implies full independence in $\mathbb{GF}(2)$ and in $\mathbb{GF}(3)$, but not in $\mathbb{GF}(P)$ with $P > 3$. The reason for this distinction is the distinction made in Lemma 1 above, regarding the possibility that a linear combination of non-uniform, independent random variables be uniform in $\mathbb{GF}(P)$ with $P > 3$ (but not in $\mathbb{GF}(2)$ or in $\mathbb{GF}(3)$).

Indeed, consider three independent random variables s_1, s_2 and s_3 in $\mathbb{GF}(5)$, with probability vectors $\boldsymbol{p}_1 = \boldsymbol{p}_2$ and $\boldsymbol{p}_3$ (resp.) following the example given in the proof of Lemma 1. Namely, let the respective characteristic vectors be given by $\tilde{\boldsymbol{p}}_1 = \tilde{\boldsymbol{p}}_2 = [1\ 0\ 0.3\ 0.3\ 0]^T$ and $\tilde{\boldsymbol{p}}_3 = [1\ 0.4\ 0\ 0\ 0.4]^T$. This implies $\boldsymbol{p}_1 = \boldsymbol{p}_2 \approx [0.32\ 0.10\ 0.24\ 0.24\ 0.10]^T$ and $\boldsymbol{p}_3 \approx [0.36\ 0.25\ 0.07\ 0.07\ 0.25]^T$. Clearly, our identifiability condition is satisfied here, since none of these random variables is uniform. However, $s_1 \oplus s_3$, as well as $s_2 \oplus s_3$, are uniform. Thus, consider the mixing-matrix $\boldsymbol{A} = \left[\begin{smallmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \end{smallmatrix}\right]$ which yields

$$\begin{aligned} x_1 &= s_1 \\ x_2 &= s_2 \\ x_3 &= s_1 \oplus s_2 \oplus s_3. \end{aligned} \tag{19}$$

Now, x_1 and x_2 are obviously statistically independent. Moreover, since $s_2 \oplus s_3$ is uniform and independent of s_1, we deduce, by Lemma 2, that x_3 and x_1 are also statistically independent. Similarly, by switching roles between s_1 and s_2, we further deduce that x_3 and x_2 are statistically independent as well. Therefore, x_1, x_2 and x_3 are pair-wise independent, but are clearly not fully mutually independent.

Obviously, such a counter-example cannot be constructed in $\mathbb{GF}(2)$ or in $\mathbb{GF}(3)$, since in these fields a linear combination of non-uniform, statistically independent random variables cannot be uniform. Furthermore, out following theorem asserts that, under our identifiability conditions, pairwise statistical independence of the mixtures indeed implies their full statistical independence in $\mathbb{GF}(2)$ and in $\mathbb{GF}(3)$.

Theorem 2: Let $\boldsymbol{s}$ be a $K \times 1$ random vector whose elements are statistically-independent, non-degenerate and non-uniform random variables in $\mathbb{GF}(2)$ or in $\mathbb{GF}(3)$. Let $\boldsymbol{y} = \boldsymbol{D} \circ \boldsymbol{s}$ denote a $K \times 1$ vector of non-trivial linear combinations of the elements of $\boldsymbol{s}$ over the field, prescribed by the elements of the $K \times K$ matrix $\boldsymbol{D}$.

If the elements of $\boldsymbol{y}$ are all pairwise statistically independent (namely, if y_k is statistically independent of y_ℓ for all $k \neq \ell$, $k, \ell \in \{1, \ldots K\}$), then $\boldsymbol{D} = \boldsymbol{\Pi} \circ \boldsymbol{\Lambda}$, where $\boldsymbol{\Pi}$ is a $K \times K$ permutation matrix and $\boldsymbol{\Lambda}$ is a $K \times K$ non-singular diagonal matrix in the field. In other words, the elements of $\boldsymbol{y}$ are merely a permutation of the (possibly scaled) elements of $\boldsymbol{s}$, and are therefore not only pairwise, but also fully statistically independent.

Obviously, in $\mathbb{GF}(2)$ $\boldsymbol{\Lambda}$ must be $\boldsymbol{I}$ (no scaling ambiguity), and in $\mathbb{GF}(3)$ (assuming the group $\{0, 1, -1\}$), $\boldsymbol{\Lambda}$ has only ± 1-s along its diagonal (the scaling ambiguity is just a sign ambiguity). A proof for each of the two cases, $\mathbb{GF}(2)$ and $\mathbb{GF}(3)$, is provided in Appendix A. We now proceed to propose practical separation approaches.

V. Practical Separation Approaches

In this section we propose two possible practical separation approaches, based on the properties developed above.

Note that any approach which exploits the full statistical description of the joint probability distribution of $\boldsymbol{x}$ would require collection (estimation) and some manipulation of the probabilities tensor $\boldsymbol{\mathcal{P}}_{\boldsymbol{x}}$, which is P^K large, and, therefore, a computational load of at least $\mathcal{O}(P^K)$ seems inevitable. Still, this is significantly smaller (and often realistically far more affordable) than $\mathcal{O}(K^2P^{(K^2)})$ (as required by brute-force search for the unmixing matrix), even for relatively small values of P and K.

Note further, that in order to obtain reasonable estimates of $\boldsymbol{\mathcal{P}}_{\boldsymbol{x}}$ in practice, the number of available observation vectors T has to be significantly larger than P^K (the size of $\boldsymbol{\mathcal{P}}_{\boldsymbol{x}}$). The estimation of $\boldsymbol{\mathcal{P}}_{\boldsymbol{x}}$ can be obtained by the following simple collection process:

1) Initialize $\widehat{\boldsymbol{\mathcal{P}}}_{\boldsymbol{x}}$ as an all-zeros tensor;
2) For $t = 1, 2, ..., T$, set $\widehat{\boldsymbol{\mathcal{P}}}_{\boldsymbol{x}}(\boldsymbol{x}[t]) \leftarrow \widehat{\boldsymbol{\mathcal{P}}}_{\boldsymbol{x}}(\boldsymbol{x}[t]) + 1$;
3) Set $\widehat{\boldsymbol{\mathcal{P}}}_{\boldsymbol{x}} \leftarrow \frac{1}{T} \cdot \widehat{\boldsymbol{\mathcal{P}}}_{\boldsymbol{x}}$.

Fortunately, however, a single collection of the observation's statistics for obtaining $\widehat{\boldsymbol{\mathcal{P}}}_{\boldsymbol{x}}$ is generally sufficient, since, in order to obtain the empirical statistical characterization $\widehat{\boldsymbol{\mathcal{P}}}_{\boldsymbol{y}}$ of any linear transformation $\boldsymbol{y} = \boldsymbol{G} \circ \boldsymbol{x}$ of the observations (where $\boldsymbol{G}$ is an arbitrary $L \times K$ matrix with elements in $\mathbb{GF}(P)$), it is not necessary to actually apply the transformation to the T available observation vectors and then recollect the probabilities. The same result can be obtained directly (without re-involving the observations), simply by applying a similar accumulation procedure to the K-way tensor $\widehat{\boldsymbol{\mathcal{P}}}_{\boldsymbol{x}}$ in constructing the L-way tensor $\widehat{\boldsymbol{\mathcal{P}}}_{\boldsymbol{y}}$:

1) Initialize $\widehat{\boldsymbol{\mathcal{P}}}_{\boldsymbol{y}}$ as an all-zeros tensor;
2) Running over all P^K index-vectors $\boldsymbol{i}$ (from $[0 \ \cdots \ 0]^T$ to $[P-1 \ \cdots \ P-1]^T$), set

$$\widehat{\boldsymbol{\mathcal{P}}}_{\boldsymbol{y}}(\boldsymbol{G} \circ \boldsymbol{i}) \leftarrow \widehat{\boldsymbol{\mathcal{P}}}_{\boldsymbol{y}}(\boldsymbol{G} \circ \boldsymbol{i}) + \widehat{\boldsymbol{\mathcal{P}}}_{\boldsymbol{x}}(\boldsymbol{i}). \tag{20}$$

Note that when $\boldsymbol{G}$ is a square invertible matrix, $\widehat{\boldsymbol{\mathcal{P}}}_{\boldsymbol{y}}$ is simply a permutation of $\widehat{\boldsymbol{\mathcal{P}}}_{\boldsymbol{x}}$.

A. Ascending Minimization of EntRopies for ICA (AMERICA)

Our first approach is based on minimizing the individual entropies of the recovered sources. Conceptually, such an approach can consist of going over all possible $P^K - 1$ nontrivial linear combinations of the observations, and computing their respective entropies. Then, given these entropies, we need to select the K linear combinations with the smallest entropies, such that their respective linear-combination coefficients vectors (rows of the implied unmixing matrix) are linearly independent (in $\mathbb{GF}(P)$).

Let us first consider the computation of the entropies of all possible (nontrivial) $P^K - 1$ linear combinations prescribed by the coefficients vectors $\boldsymbol{i}_n$ (for $n = 1, ..., P^K - 1$). Each requires the computation of the respective probabilities vector $\boldsymbol{p}_{y_n}$ of $y_n = \boldsymbol{i}_n^T \circ \boldsymbol{x}$, by applying the above-mentioned tensor-accumulation procedure with $\boldsymbol{G} = \boldsymbol{i}_n^T$ to the tensor $\widehat{\boldsymbol{\mathcal{P}}}_{\boldsymbol{x}}$. Thus, the number of required multiplications is roughly $\mathcal{O}(K \cdot (P^K)^2) = \mathcal{O}(K \cdot P^{2K})$, which (for $K > 2$) is much smaller than $\mathcal{O}(K^2 \cdot P^{(K^2)})$ (the brute-force search cost), but may still be quite large. Fortunately, it is possible to compute the required probabilities vectors more conveniently, via the estimated characteristic tensor $\widehat{\widetilde{\boldsymbol{\mathcal{P}}}}_{\boldsymbol{x}}$, which can be obtained using a multidimensional Fast Fourier Transform (FFT).

The proposed computation proceeds as follows. First, given the estimated probabilities tensor $\widehat{\boldsymbol{\mathcal{P}}}_{\boldsymbol{x}}$, we obtain the estimated characteristic tensor $\widehat{\widetilde{\boldsymbol{\mathcal{P}}}}_{\boldsymbol{x}}$ using a K-dimensional FFT, by successively applying 1-dimensional radix-P DFTs along each of the K dimensions. Thus, for each dimension we compute P^{K-1} P-long DFTs, at the cost of $\mathcal{O}(P^{K-1} \cdot (P \log P)) = \mathcal{O}(P^K \log P)$. The total cost for obtaining $\widehat{\widetilde{\boldsymbol{\mathcal{P}}}}_{\boldsymbol{x}}$ is therefore $\mathcal{O}(K \cdot P^K \log P) = \mathcal{O}(P^K \log(P^K))$, rather than $\mathcal{O}((P^K)^2)$, as would be required by direct calculation.

Now, in order to obtain the characteristic vector $\tilde{\boldsymbol{p}}_{y_n}$ of $y_n = \boldsymbol{i}_n^T \circ \boldsymbol{x}$, we can exploit the following relation:

$$\tilde{p}_{y_n}(m) = E[W_P^{m y_n}] = E[W_P^{m \boldsymbol{i}_n^T \boldsymbol{x}}] = \widetilde{\mathcal{P}}_{\boldsymbol{x}}(m \otimes \boldsymbol{i}_n), \quad m = 0, ..., P-1, \tag{21}$$

which means that for each $\boldsymbol{i}_n$, each (m-th) element of the characteristic vector of y_n can be extracted from the respective element ($m \otimes \boldsymbol{i}_n$) of $\widetilde{\boldsymbol{\mathcal{P}}}_{\boldsymbol{x}}$. Note further, that the first ($m = 0$) element of each characteristic vector is 1; and that the conjugate-symmetry of the characteristic vectors can be exploited, such that only the ”first half” ($m = 1, ..., \lfloor P/2 \rfloor$) needs to be extracted from $\widetilde{\boldsymbol{\mathcal{P}}}_{\boldsymbol{x}}$. Naturally, in the absence of the true $\widetilde{\boldsymbol{\mathcal{P}}}_{\boldsymbol{x}}$, we would use the empirical $\widehat{\widetilde{\boldsymbol{\mathcal{P}}}}_{\boldsymbol{x}}$, obtained from the empirical probabilities tensor $\widehat{\boldsymbol{\mathcal{P}}}_{\boldsymbol{x}}$, as described above.

The extraction of the characteristic vectors $\tilde{\boldsymbol{p}}_{y_n}$ for all $\boldsymbol{i}_n$ requires $\mathcal{O}(P^K \cdot PK)$ additional operations. Once these vectors are obtained, they are each converted, using inverse FFT, into probabilities vectors $\boldsymbol{p}_{y_n}$, from which the entropies are readily obtained. This requires additional $\mathcal{O}(P^K \cdot (P \log P + P))$ operations (excluding the computation of $P \cdot P^K$ logarithms).

Given the entropies of all possible linear combinations (ignoring the trivial $\boldsymbol{i}_0 = \boldsymbol{0}$), the one with the smallest entropy corresponds to the first extracted source. Once the smallest-entropy source is identified, a ”natural” choice is to proceed to the linear combination yielding the second-smallest entropy (and so forth), but special care has to be taken, so that each selected coefficients vectors should not be linearly dependent (in $\mathbb{GF}(P)$) on the previous ones. One possible way to assure this, is to take a “deflation” approach (also sometimes taken in classical ICA - see, e.g., [9] or [1]), in which each extracted source is first eliminated from the mixture, and then the lowest-entropy combination of the remaining (“deflated”) mixtures is taken as the “next” extracted source. However, such an approach requires finding the coefficients needed for elimination of the extracted source from each mixture element, as well as recalculation of all the entropies after each deflation stage, which seems computationally expensive. A possible alternative is to use a greedy sequential extraction, such that the k-th chosen coefficients vector is the one associated with the smallest entropy while being linearly independent of the previously selected $k-1$ coefficients vectors. Checking whether a $K \times 1$ vector $\hat{\boldsymbol{b}}_k$ is linearly independent of the $K \times 1$ vectors $\hat{\boldsymbol{b}}_1, \hat{\boldsymbol{b}}_2, ..., \hat{\boldsymbol{b}}_{k-1}$ amounts to checking whether there exists a nonzero $k \times 1$ vector $\boldsymbol{\alpha}$, such that $[\hat{\boldsymbol{b}}_1 \;\; \cdots \;\; \hat{\boldsymbol{b}}_k] \circ \boldsymbol{\alpha} = \mathbf{0}$, which can be checked by an exhaustive search among all possible nonzero $k \times 1$ vectors in $\mathbb{GF}(P)$. This roughly adds (in the "worst", last stage, with $k = K$) $\mathcal{O}(K^2 \cdot P^K)$ multiplications.

The total computational cost is therefore approximately $\mathcal{O}(P^K \cdot (K^2 + KP + K \log P + P \log P + P))$. The proposed algorithm, which was given the acronym "AMERICA" (Ascending Minimization of EntRopies for ICA) is summarized in Table 1.

```
Algorithm 1: AMERICA
Input: $\widehat{\boldsymbol{\mathcal{P}}}_{\boldsymbol{x}}$ - the mixtures' $K$-way $P \times P \times \cdots \times P$ estimated (empirical) probabilities tensor;
Output: $\hat{\boldsymbol{B}}$ - the $K \times K$ estimated separation matrix;
Notations: We denote by the $K \times 1$ $P$-nary vector $\boldsymbol{i}_n$ the $n$-th index vector (for $n = 0, ..., P^K - 1$), such that $n = \sum_{k=1}^{K} i_n(k) P^{k-1}$, where $\boldsymbol{i}_n = [i_n(1) \ \cdots \ i_n(K)]^T$; All indices in the description below run from $0$.
Algorithm:
 1) Compute $\hat{\widehat{\boldsymbol{\mathcal{P}}}}_{\boldsymbol{x}}$, the observations' empirical characteristic tensor, by applying a $K$-dimensional radix-$P$ FFT to $\widehat{\boldsymbol{\mathcal{P}}}_{\boldsymbol{x}}$.
 2) For $n = 0, ..., P^K - 1$, compute $h_n$, the (empirical) entropy of the random variable $y_n \stackrel{\triangle}{=} \boldsymbol{i}_n^T \circ \boldsymbol{x}$ as follows:
    a) Obtain the $P \times 1$ empirical characteristic vector of $y_n$, denoted $\tilde{\boldsymbol{p}}_n$, as follows:
         i) Set $\tilde{p}_n(0) := 1$;
        ii) Set $\tilde{p}_n(1) := \hat{\widehat{\mathcal{P}}}_{\boldsymbol{x}}(\boldsymbol{i}_n)$;
       iii) If $P = 3$, set $\tilde{p}_n(2) := \hat{\widehat{\mathcal{P}}}^*_{\boldsymbol{x}}(\boldsymbol{i}_n)$;
        iv) If $P > 3$, then for $m = 2, ..., (P-1)/2$, set $\tilde{p}_n(m) := \hat{\widehat{\mathcal{P}}}_{\boldsymbol{x}}(m \otimes \boldsymbol{i}_n)$ and $\tilde{p}_n(P+1-m) := \hat{\widehat{\mathcal{P}}}^*_{\boldsymbol{x}}(m \otimes \boldsymbol{i}_n)$;
    b) Obtain the $P \times 1$ empirical probabilities vector of $y_n$, denoted $\boldsymbol{p}_n$, by applying an inverse FFT to the vector $\tilde{\boldsymbol{p}}_n$;
    c) Obtain $h_n = \sum_{m=0}^{P-1} p_n(m) \log p_n(m)$;
 3) Find the smallest entropy among $h_1, ..., h_{P^K-1}$ and denote the minimizing index $n_1$ (i.e., $h_{n_1} = \min_{n \neq 0} h_n$);
 4) Set $\hat{\boldsymbol{B}} := \boldsymbol{i}_{n_1}^T$ and mark $h_{n_1}$ as "used";
 5) Repeat for $k = 2, ..., K$:
    a) Find the smallest among all "unused" entropies; denote the minimizing index $n_k$;
    b) Construct the test-matrix $\bar{\boldsymbol{B}} := [\hat{\boldsymbol{B}}^T \ \boldsymbol{i}_{n_k}]$;
    c) Go over all nonzero length-$k$ index vectors $\boldsymbol{j}_n$ ($n = 1, ..., p^k - 1$), checking whether $\bar{\boldsymbol{B}} \circ \boldsymbol{j}_n = \boldsymbol{0}$ for some $n$. If such $\boldsymbol{j}_n$ is found, mark $h_{n_k}$ as "used" and find the next smaller entropy (i.e., go to step 5a);
    d) Set $\hat{\boldsymbol{B}} := \bar{\boldsymbol{B}}^T$.
```

B. Minimizing Entropies by eXchanging In COuples (MEXICO)

An alternative separation approach, which avoids prior calculation of the entropies of all possible linear combinations, is to try to find the separating transformation by successively minimizing the entropies in couples (going over all couples combinations in each "sweep"). More specifically, let x_1 and x_2 denote the first two elements of the mixtures vector, and let $\boldsymbol{P}_{1,2}$ denote their $P \times P$ joint probability matrix, which can be obtained from the tensor $\boldsymbol{\mathcal{P}}_{\boldsymbol{x}}$ by summing along all other dimensions:

$$\boldsymbol{P}_{1,2}(m,n) = \sum_{i_3,\ldots,i_K=0}^{P-1} \mathcal{P}_{\boldsymbol{x}}(m,n,i_3,\ldots,i_K) \quad m,n \in [0,P-1]. \tag{22}$$

Consider a random variable of the form

$$\bar{x}_1 = x_1 \oplus c \otimes x_2, \tag{23}$$

where $c \in [1, P-1]$ is some constant. Let $\boldsymbol{p}_{\bar{x}_1}(c)$ denote the probabilities vector of $\bar{x}_1$. The m-th element of this vector is given (depending on c) by

$$p_{\bar{x}_1}(m;c) = \Pr\{x_1 \oplus c \otimes x_2 = m\} = \sum_{n=0}^{P-1} \Pr\{x_1 = n, c \otimes x_2 = m \ominus n\} = \sum_{n=0}^{P-1} \boldsymbol{P}_{1,2}(n, c^{-1} \otimes (m \ominus n))\}, \tag{24}$$

where c^{-1} denotes the reciprocal of c in $\mathbb{GF}(P)$, such that $c \otimes c^{-1} = 1$. The entropy of $\bar{x}_1$ is then given by

$$H(\bar{x}_1; c) = -\sum_{m=0}^{P-1} p_{\bar{x}_1}(m;c) \log p_{\bar{x}_1}(m;c). \tag{25}$$

COnsider the value c_0 of c which minimizes $H(\bar{x}_1; c)$. If the resulting entropy is smaller than the entropy of x_1, then substitution of x_1 with $\bar{x}_1 = x_1 \oplus c_0 \otimes x_2$ in $\boldsymbol{x}$ would be an invertible linear transformation which reduces the sum of entropies of the elements of $\boldsymbol{x}$.

Note ,in addition, that following this transformation the mutual information $I(\bar{x}_1, x_2) = H(\bar{x}_1) + H(x_2) - H(\bar{x}_1, x_2)$ will be smaller than $I(x_1, x_2)$, because the joint entropies $H(x_1, x_2)$ and $H(\bar{x}_1, x_2)$ are the same (since the transformation is invertible). Therefore, this transformation also makes these two elements "more independent".

Thus, based on this basic operation, a separation approach can be taken as follows. Let $\boldsymbol{y}$ denote the random vector of "demixed" sources to be constructed by successive linear transformations of $\boldsymbol{x}$, and initialize $\boldsymbol{y} = \boldsymbol{x}$, along with its probabilities tensor $\boldsymbol{\mathcal{P}}_{\boldsymbol{y}} = \boldsymbol{\mathcal{P}}_{\boldsymbol{x}}$. Proceed sequentially through all couples y_k, y_ℓ in $\boldsymbol{y}$: For each couple, compute the joint probabilities matrix $\boldsymbol{P}_{k,\ell}$, and then look for the value of c which minimizes the entropy of $\bar{y}_k = y_k \oplus c \otimes y_\ell$. If this entropy is smaller than that of y_k, replace y_k with $\bar{y}_k$, recording the implied linear transformation as $\bar{\boldsymbol{y}} = \boldsymbol{V}(k,\ell;c) \circ \boldsymbol{y}$, where

$$\boldsymbol{V}(k,\ell;c) \stackrel{\triangle}{=} \boldsymbol{I} + c \cdot \boldsymbol{E}_{k,\ell}, \tag{26}$$

$\boldsymbol{E}_{k,\ell}$ denoting a $K \times K$ all-zeros matrix with a 1 at the (k,ℓ)-th position. If the minimal entropy of $\bar{y}_k$ is larger than that of y_k, no update takes place, and the next couple is addressed.

Upon an update, $\bar{\boldsymbol{y}}$ serves as the new $\boldsymbol{y}$. The probabilities tensor $\boldsymbol{\mathcal{P}}_{\boldsymbol{y}}$ is updated accordingly (this update is merely a permutation, attainable using (20) with $\boldsymbol{G} = \boldsymbol{V}(k,\ell;c)$). The procedure is repeated for each indices-couple (k,ℓ) (with $k \neq \ell$), and we term a "sweep" as a sequential pass over all possible $K(K-1)$ combinations (note that there is no symmetry here, namely, the couple (ℓ,k) is essentially different from (k,ℓ)). Sweeps are repeated sequentially, until a full seep without a single update occurs, which terminates the process.

In practice, the algorithm is applied starting with the empirical observations' probabilities tensor $\widehat{\boldsymbol{\mathcal{P}}}_{\boldsymbol{x}}$, and the accumulated sequential left-product of the $\boldsymbol{V}(k,\ell;c)$ matrices yields the estimated separating matrix. Since the sum of marginal entropies of the elements of $\boldsymbol{y}$ is bounded below and is guaranteed not to increase (usually to decrease) in each sweep, and since the algorithm stops upon encountering the first sweep without such a decrease - such a stop is guaranteed to occur within a finite number of sweeps.

Note, however, that in general there is no guarantee for consistent separation using this algorithm, i.e., even if the true probabilities tensor $\boldsymbol{\mathcal{P}}_{\boldsymbol{x}}$ of the observations is known (and used), the stopping point is generally not guaranteed to imply separation. The rationale behind this algorithm is the hope that such a "pairwise separation" scheme would ultimately yield pairwise independence, which, at least for $P = 2$ and $P = 3$, would in turn imply full independence (hence separation), per Theorem 2 above. Strictly speaking, however, this algorithm is not even guaranteed to yield pairwise separation. For example, consider the $P = 2$ case, with a mixing matrix

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}, \tag{27}$$

when all the sources have equal $p(1)$ (probability of taking the value 1). In this particular case, the number of 1-s in a linear combination of any two lines is greater or equal to the number of 1-s in each of the two lines. Therefore, there is no pairwise linear combination which reduces the entropy of any of the mixtures in this case. Therefore, the algorithm may stop short of full separation when such a condition is encountered.

Nevertheless, such conditions are relatively rare, and, as we show in simulation results in the following section, this algorithm is quite successful. Its leading advantage over AMERICA is in its reduced computational complexity when the unmixing matrix $\boldsymbol{B}$ is sparse and $K >> P$.

Indeed, the computational complexity of this iterative algorithm naturally depends on the number of required sweeps and on the number of updates in each sweeps - which in turn depend strongly on the true mixing matrix $\boldsymbol{A}$ (and, to some extent, also on sources' realizations). Testing each couple (k, ℓ) requires computation of the joint probabilities matrix $\boldsymbol{P}_{k,\ell}$ - which requires $\mathcal{O}(P^K)$ additions (no multiplications are needed). Then, looking for the optimal c requires $P - 1$ computations of the probabilities vector of the respective $\bar{y}_k$ - a total of additional $\mathcal{O}(P^3)$ additions (again, no multiplications are needed for this) and $\mathcal{O}(P^2)$ log operations. If an update takes place, recalculation of $\widehat{\boldsymbol{\mathcal{P}}}_{\boldsymbol{y}}$ is also needed, which is $\mathcal{O}(P^K)$ (but, as mentioned above, this is merely a permutation of the tensor).

Therefore, the first sweep requires $\mathcal{O}(P^2(P^K + P^3)) = \mathcal{O}(P^{K+2} + P^5))$ operations and $\mathcal{O}(P^4)$ log operations, plus $\mathcal{O}(P^K)$ for each update within the sweep. Naturally, a couple tested in one sweep does not have to be tested in a subsequent sweep if no substitution involving any of its members had occurred in the former. Therefore, for subsequent sweeps the number of operations can be significantly smaller, depending on the number of updates occurring along the way - which is obviously data-dependent. The number of required sweeps is also data dependent.

Thus, the computational complexity of this algorithm, assuming $K > 3$, can be roughly estimated at $\mathcal{O}(P^K \cdot (N_d P^2))$, where N_d denotes a data-dependent constant, which can be very small (of the order of $2-3$) when the true demixing matrix $\boldsymbol{B}$ is very sparse (only a few sweeps with few updates are needed), but can be considerably large when $\boldsymbol{B}$ is rather "rich". Compared to the computational complexity of AMERICA, we observe that, assuming $K >> P$, this algorithm is preferable if $N_d P^2 < K^2$.

The algorithm, which was given the acronym "MEXICO" (Minimizing Entropies by eXchanging In COuples) is summarized in Table 2.

Algorithm 2: MEXICO

```
Input: $\widehat{\boldsymbol{\mathcal{P}}}_{\boldsymbol{x}}$ - the mixtures' $K$-way $P \times P \times \cdots \times P$ estimated (empirical) probabilities tensor;
Output: $\hat{\boldsymbol{B}}$ - the $K \times K$ estimated separation matrix;
Algorithm:
 1) Initialize: $\hat{\boldsymbol{B}} := \boldsymbol{I}$. Conceptually, we denote the "demixed" random vector $\boldsymbol{y} \stackrel{\triangle}{=} \hat{\boldsymbol{B}} \circ \boldsymbol{x}$, so set $\widehat{\boldsymbol{\mathcal{P}}}_{\boldsymbol{y}} := \widehat{\boldsymbol{\mathcal{P}}}_{\boldsymbol{x}}$;
 2) Initialize: $\boldsymbol{h} = [h_1 \cdots h_K]^T$ with the empirical entropies of the $K$ respective elements $y_k$ (each computed from the empirical probabilities vector, which is obtained by summation over all other ($\neq k$) dimensions in $\widehat{\boldsymbol{\mathcal{P}}}_{\boldsymbol{y}}$);
 3) Initialize $\boldsymbol{F}$, as a $K \times K$ all-ones flags matrix: $F(k,\ell) = 1$ means that the $(k,\ell)$-th couple needs to be (re)tested;
 4) Run a "sweep": Repeat for $k = 1, \ldots, K$, for $\ell = 1, \ldots, K$, $\ell \neq k$
    If $F(k,\ell) = 1$ do the following:
   a) Compute $\boldsymbol{P}_{k,\ell}$, the empirical joint probabilities matrix of $y_k$ and $y_\ell$, by summation over all other dimensions ($\neq k, \ell$) in $\widehat{\boldsymbol{\mathcal{P}}}_{\boldsymbol{y}}$;
   b) For $c = 1, \ldots, P-1$, compute the elements of $\boldsymbol{p}_{\bar{y}_k}(c)$, the probabilities vector of $\bar{y}_k = y_k \oplus c \otimes y_\ell$, in a way similar to (24), yielding its entropy $H(\bar{y}_k; c)$;
   c) Denote the minimum entropy as $H_0 = H(\bar{y}_k; c_0)$ (with $c_0$ denoting the minimizing $c$);
   d) If $H_0 < h_k$ apply a substitution:
       i) Set $\boldsymbol{V} = \boldsymbol{I} + c_0 \cdot \boldsymbol{E}_{k,\ell}$;
      ii) Update $\hat{\boldsymbol{B}} := \boldsymbol{V} \circ \hat{\boldsymbol{B}}$;
     iii) Update the probabilities tensor using (20) with $\boldsymbol{G} = \boldsymbol{V}$;
      iv) Mark all couples involving $k$ as "need to be retested":
          $\boldsymbol{F}(k,:) := 1$, $\boldsymbol{F}(:,k) := 1$;
       v) Update $h_k := H_0$;
      vi) (Conceptually: $\boldsymbol{y} := \boldsymbol{V} \circ \boldsymbol{y}$);
   e) Mark the $(k,\ell)$-th element as "tested": $F(k,\ell) = 0$, and proceed;
 5) If $\boldsymbol{F} \neq \boldsymbol{I}$ (there are still couples to be (re)tested), run another sweep; Else stop.
```

VI. Rudimentary performance analysis and simulation results

In this section we present a rudimentary analysis of the expected performance of the proposed algorithms, in order to obtain an estimate of the expected rate of success in separating the sources, at least in some simple cases.

Let us first establish the concept of *equivariance*. In classical ICA, an algorithm is called equivariant (see, e.g., [10]) with respect to the mixing matrix $\boldsymbol{A}$, if its performance does not depend on $\boldsymbol{A}$ (as long as it is invertible), but only on the realization of the sources. This appealing property is shared by many (but certainly not by all) classical ICA algorithms (in the context of noiseless classical ICA).

We shall now show that, with some slight modification, the AMERICA algorithm is equivariant. Recall that AMERICA is based on computation of all the empirical probabilities vectors $\boldsymbol{p}_{y_n}$ of the random variables $y_n = \boldsymbol{i}_n^T \circ \boldsymbol{x}$ for all possible index-combinations $\boldsymbol{i}_n$, followed by sequential extraction of the index-vectors $\boldsymbol{i}_n$ corresponding to the smallest entropies (while maintaining sequential mutual linear independence). Although not directly calculated in this way in the algorithm, the ℓ-th element of $\boldsymbol{p}_{y_n}$ is evidently given by

$$p_{y_n}(\ell) = \widehat{\Pr}\{\boldsymbol{i}_n^T \circ \boldsymbol{x} = \ell\} = \frac{1}{T}\sum_{t=1}^{T} I\{\boldsymbol{i}_n^T \circ \boldsymbol{x}[t] = \ell\}, \tag{28}$$

where $\widehat{\Pr}\{\cdot\}$ denoted the empirical probability, and where $I\{\cdot\}$ denotes the Indicator function. But since $\boldsymbol{x}[t] = \boldsymbol{A} \circ \boldsymbol{s}[t]$, we obviously have

$$I\{\boldsymbol{i}_n^T \circ \boldsymbol{x}[t] = \ell\} = I\{(\boldsymbol{A}^T \circ \boldsymbol{i}_n)^T \circ \boldsymbol{s}[t] = \ell\}, \tag{29}$$

which means that with any given realization $\boldsymbol{s}[1], \cdots, \boldsymbol{s}[T]$ of the sources, the empirical probabilities vector $\boldsymbol{p}_{y_n}$ of $y_n = \boldsymbol{i}_n^T \circ \boldsymbol{x}$ obtained when the mixing matrix is $\boldsymbol{A}$, is equal to some empirical probabilities vector $\boldsymbol{p}_{y_m}$ of $y_m = \boldsymbol{i}_m^T \circ \boldsymbol{x}$ obtained when the mixing matrix is $\boldsymbol{I}$ (i.e., when there is no mixing), such that $\boldsymbol{i}_m = \boldsymbol{A}^T \boldsymbol{i}_n$. Since $\boldsymbol{A}$ is invertible, this relation is bijective, which implies that the $P^K - 1$ empirical probabilities vectors obtained with any (invertible) mixing are merely a permutation of the same set of $P^K - 1$ vectors that would be obtained when the sources are not mixed. Consequently, if, based on the empirical entropies of these empirical vectors, the matrix $\hat{\boldsymbol{B}} = [\boldsymbol{i}_{n_1}\ \boldsymbol{i}_{n_2}\ \cdots\ \boldsymbol{i}_{n_K}]^T$ is formed by the algorithm when the mixing-matrix is $\boldsymbol{A}$, this implies that the matrix

$$\hat{\boldsymbol{B}}_0 \stackrel{\triangle}{=} [\boldsymbol{i}_{m_1}\ \boldsymbol{i}_{m_2}\ \cdots\ \boldsymbol{i}_{m_K}]^T = \left(\boldsymbol{A}^T \circ [\boldsymbol{i}_{n_1}\ \boldsymbol{i}_{n_2}\ \cdots\ \boldsymbol{i}_{n_K}]\right)^T = \hat{\boldsymbol{B}} \circ \boldsymbol{A} \tag{30}$$

would be formed by the algorithm when the sources are unmixed. Consequently, the overall mixing-unmixing matrix[2] $\hat{\boldsymbol{B}} \circ \boldsymbol{A}$ in the mixed case would equal the overall mixing-unmixing matrix $\hat{\boldsymbol{B}}_0 \circ \boldsymbol{I} = \hat{\boldsymbol{B}} \circ \boldsymbol{A}$ in the unmixed case. This means that, no matter what the (invertible) mixing matrix is, the

[2] This matrix is sometimes also called the “contamination” matrix, describing the residual mixing (if any).

overall mixing-unmixing matrix would be the same as would be obtained by the AMERICA algorithm in the unmixed case - implying the desired equivariance property.

There is, however, one small caveat that has to be considered. The reasoning above assumes that the sequential progress of the algorithm through the sorted empirical entropies for selecting, testing (for linear dependence) and using the index-vectors is uniquely determined by the calculated entropy values, and is independent of the values of the index-vectors. This is generally true, with one possible exception: If the set of empirical entropies happens to contain a subset with equal entropies, the (arbitrary) order in which the index-vectors within such a subset are sorted is usually lexicographic - which introduces dependence on the actual index values, and such dependence is not permutation-invariant - thereby potentially introducing dependence on the mixing matrix in turn. In order to avoid this condition, any sub-group with equal empirical entropies should be somehow inner-sorted in a way which is independent of corresponding index-vectors values - e.g., by randomization. Note that the occurrence of such a subset (with empirical entropies that are exactly equal) becomes very rare when the number of observations T is large, but may certainly happen when T is relatively small. Note further, that with such randomization the attained separation for a given realization depends not only on the sources' realization, but also on this random sorting within subsets (but not on the mixing matrix), and therefore only statistical measures of the performance (e.g., the probability of perfect separation) can be considered equivariant.

Having established the equivariance, we now proceed to analyze the probability of perfect separation in the most simple case: $P = 2$, $K = 2$. Thanks to the equivariance property we may assume, without loss of generality, that the mixing matrix is the identity matrix, $\boldsymbol{A} = \boldsymbol{I}$. Let $p_{s_1}(1) = \rho_1$ (resp., $p_{s_2}(1) = \rho_2$) denote the probability with which the first (resp., second) source takes the value 1. Due to the assumed non-mixing conditions ($\boldsymbol{A} = \boldsymbol{I}$), these are also the probabilities of the "mixtures" x_1 and x_2. To characterize the empirical probabilities tensor $\widehat{\boldsymbol{\mathcal{P}}}_{\boldsymbol{x}}$, let us denote by N_{00}, N_{01}, N_{10} and N_{11} the number of occurrences of $\boldsymbol{x}[t] = [0\ 0]^T$, $\boldsymbol{x}[t] = [0\ 1]^T$, $\boldsymbol{x}[t] = [1\ 0]^T$ and $\boldsymbol{x}[t] = [1\ 1]^T$ (resp.) within the observed sequence of length T. Thus, the elements of the 2×2 empirical probabilities tensor (matrix in this case) are $\widehat{\mathcal{P}}_{\boldsymbol{x}}(m_1, m_2) = N_{m_1,m_2}/T$, for $m_1, m_2 \in \{0, 1\}$.

The empirical probability $\hat{p}_{x_1}(1)$ of x_1 taking the value 1 is given by $\widehat{\mathcal{P}}_{\boldsymbol{x}}(1,0) + \widehat{\mathcal{P}}_{\boldsymbol{x}}(1,1) = (N_{10} + N_{11})/T$. The empirical probability $\hat{p}_{x_1 \oplus x_2}(1)$ of the random variable $x_1 \oplus x_2$ taking the value 1 is given by $\widehat{\mathcal{P}}_{\boldsymbol{x}}(1,0) + \widehat{\mathcal{P}}_{\boldsymbol{x}}(0,1) = (N_{10} + N_{01})/T$. An identification error would occur if the entropy associated with the latter be smaller than that associate with the former (because then the (wrong) linear combination vector $\boldsymbol{i}_3^T = [1\ 1]$ would be preferred by the algorithm over the (correct) linear combination vector $\boldsymbol{i}_1^T = [1\ 0]$ as a row in $\hat{\boldsymbol{B}}$).

In the $P = 2$ case, the entropy is monotonically decreasing in the distance of $p(1)$ (or $p(0)$) from $\frac{1}{2}$. Assuming that T is "sufficiently large", the empirical $\hat{p}_{x_1}(1)$ would be close to its true value ρ_1, and the empirical $\hat{p}_{x_1 \oplus x_2}(1)$ would be close to its true value $\rho_1(1-\rho_2)+\rho_2(1-\rho_1) = \rho_1+\rho_2-2\rho_1\rho_2$. Assuming that $\rho_1, \rho_2 < \frac{1}{2}$, both ρ_1 and $\rho_1 + \rho_2 - 2\rho_1\rho_2$ are smaller than $\frac{1}{2}$, and we can therefore assume that so are the empirical $\hat{p}_{x_1}(1)$ and $\hat{p}_{x_1 \oplus x_2}(1)$. Thus, the empirical entropy associated with the linear combination $x_1 \oplus x_2$ would be smaller than that associated with x_1 if

$$\hat{p}_{x_1 \oplus x_2}(1) < \hat{p}_{x_1}(1) \quad \Leftrightarrow \quad \tfrac{1}{T}(N_{10} + N_{01}) < \tfrac{1}{T}(N_{10} + N_{11}) \quad \Leftrightarrow \quad N_{01} < N_{11}. \tag{31}$$

We are therefore interested in the probability of the event $\Xi 1 : \;\; N_{01} < N_{11}$. Let us denote by $N_2 \overset{\triangle}{=} N_{01} + N_{11}$ the number of occurrences of $x_2[t] = 1$ in $[1, T]$. The probability of $\Xi 1$ can then be expressed as follows:

$$\begin{aligned}\Pr\{\Xi 1\} = \Pr\{N_{01} < N_{11}\} = \Pr\{N_{11} > \tfrac{1}{2}N_2\} = \sum_{M=1}^{T} \Pr\{N_2 = M \;\cap\; N_{11} > \tfrac{1}{2}M\} = \\ \sum_{M=1}^{T} \Pr\{N_2 = M\} \Pr\{N_{11} > \tfrac{1}{2}M | N_2 = M\}.\end{aligned} \tag{32}$$

Due to the statistical independence between the sources (and therefore between x_1 and x_2), given that $N_2 = M$, the random variable N_{11} is simply the number of occurrences of $x_1[t] = 1$ among M independent trials - a Binomial random variable with M trials and probability ρ_1, which we shall denote as $N_{1,M} \sim B(M, \rho_1)$. Thus,

$$\begin{aligned}\Pr\{\Xi 1\} = \Pr\{N_{01} < N_{11}\} = \sum_{M=1}^{T} \Pr\{N_2 = M\} \Pr\{N_{1,M} > \tfrac{1}{2}M\} = \\ \sum_{M=1}^{T} \binom{T}{M} \rho_2^M (1-\rho_2)^{T-M} \cdot \sum_{N=\left\lfloor \frac{M}{2} \right\rfloor + 1}^{M} \binom{M}{N} \rho_1^N (1-\rho_1)^{M-N}.\end{aligned} \tag{33}$$

The inner sum is the complementary cumulative distribution function of the binomial distribution, which can also be expressed using the *normalized incomplete beta function*[3],

$$\begin{aligned}\Pr\{N_{1,M} > \tfrac{1}{2}M\} = 1 - \Pr\{N_{1,M} \le \tfrac{1}{2}M\} = \\ 1 - I_{1-\rho_1}(M - \lfloor \tfrac{1}{2}M \rfloor, \lfloor \tfrac{1}{2}M \rfloor + 1) = I_{\rho_1}(\lfloor \tfrac{1}{2}M \rfloor + 1, \lceil \tfrac{1}{2}M \rceil),\end{aligned} \tag{34}$$

with

$$I_p(n, m) \overset{\triangle}{=} n \binom{n+m-1}{m-1} \int_0^p t^{n-1}(1-t)^{m-1} dt = 1 - I_{1-p}(m, n). \tag{35}$$

Note further (from (33)), that the probability of $\Xi 1$ can be expressed as

$$\Pr\{\Xi 1\} = E\left[I_{\rho_1}(\lfloor \tfrac{1}{2}N_2 \rfloor + 1, \lceil \tfrac{1}{2}N_2 \rceil)\right] \tag{36}$$

[3] See, e.g., Binomial Distribution from Wikipedia [online], available: `http://en.wikipedia.org/wiki/Binomial_distribution`

(where the expectation is taken with respect to N_2). When $\rho_2 \cdot T$ is ”sufficiently large” this probability may be approximated by substituting N_2 with its mean, $E[N_2] = \rho_2 \cdot T$,

$$\Pr\{\Xi 1\} \approx I_{\rho_1}(\lfloor \tfrac{\rho_2}{2} T \rfloor + 1, \lceil \tfrac{\rho_2}{2} T \rceil). \tag{37}$$

The event $\Xi 1$ is just one possible component of an error event in which the algorithm would prefer the (wrong) linear combination vector $\boldsymbol{i}_3^T = [1\ 1]$ over the (correct) linear combination vector $\boldsymbol{i}_1^T = [1\ 0]$. Such an error may also happen when the empirical entropies of x_1 and of $x_1 \oplus x_2$ are equal, namely when $N_{01} = N_{11}$: assuming that the algorithm makes a random decision in such cases (to ensure mean equivariance, as discussed above), the probability of an error being caused by this event (denoted $\Xi 2$) would be $\frac{1}{2}\Pr\{\Xi 2\}$. Evidently,

$$\Pr\{\Xi 2\} = \Pr\{N_{01} = N_{11}\} = \sum_{M=0}^{T} \Pr\{N_2 = M\} \Pr\{N_{1,M} = \tfrac{1}{2} M\} =$$

$$\sum_{M'=0}^{\lfloor T/2 \rfloor} \tbinom{T}{2M'} \rho_2^{2M'} (1-\rho_2)^{T-2M'} \cdot \tbinom{2M'}{M'} \rho_1^{M'} (1-\rho_1)^{M'} = \sum_{M=0}^{\lfloor T/2 \rfloor} \tfrac{T!(1-\rho_2)^T}{(T-2M)!(M!)^2} \left(\tfrac{\rho_2^2 \rho_1 (1-\rho_1)}{(1-\rho_2)^2} \right)^M . \tag{38}$$

Note that since the event $N_{1,M} = \frac{1}{2} M$ can only happen for even values of M, an approximation using the mean with respect to N_2 (as used for $\Pr\{\Xi 1\}$ above) would be far less accurate, and would therefore not be pursued.

Summarizing this part of the error analysis, the probability that the algorithm would wrongly prefer $\boldsymbol{i}_3^T = [1\ 1]$ over $\boldsymbol{i}_1^T = [1\ 0]$ as a row in $\hat{\boldsymbol{B}}$ can be approximated as

$$\Pr\{\Xi 1\} + \tfrac{1}{2}\Pr\{\Xi 2\} \approx I_{\rho_1}(\lfloor \tfrac{\rho_2}{2} T \rfloor + 1, \lceil \tfrac{\rho_2}{2} T \rceil) + \tfrac{1}{2} \cdot \sum_{M=0}^{\lfloor T/2 \rfloor} \tfrac{T!(1-\rho_2)^T}{(T-2M)!(M!)^2} \left(\tfrac{\rho_2^2 \rho_1 (1-\rho_1)}{(1-\rho_2)^2} \right)^M . \tag{39}$$

An error in the “opposite” direction occurs when the algorithm prefers $\boldsymbol{i}_3^T = [1\ 1]$ over $\boldsymbol{i}_2^T = [0\ 1]$ as a row in $\hat{\boldsymbol{B}}$. The probability of this kind of error is evidently given by the same expressions by swapping the roles of ρ_1 and ρ_2. A failure of the algorithm is defined as the occurrence of either one of the two errors. Although they are certainly not mutually exclusive, we can still approximate (or at least provide an approximate upper-bound for) the probability of occurrence of either one, by the sum of probabilities of occurrence of each. Assuming, for further simplicity of the exposition, that $\rho_1 = \rho_2 = \rho$, the approximate probability of failure is given by

$$\Pr\{\text{Failure}\} \approx 2 \cdot I_\rho(\lfloor \tfrac{\rho}{2} T \rfloor + 1, \lceil \tfrac{\rho}{2} T \rceil) + \sum_{M=0}^{\lfloor T/2 \rfloor} \tfrac{T!(1-\rho)^T}{(T-2M)!(M!)^2} \left(\tfrac{\rho^3}{1-\rho} \right)^M . \tag{40}$$

Recall that two assumptions are necessary for this approximation to hold: i) that ρ is sufficiently smaller than 0.5; and ii) that $\rho \cdot T$ is sufficiently large.

In order to test this approximation we simulated the mixing and separation of $K = 2$ independent binary ($P = 2$) sources, each taking the value 1 with probability ρ. In Fig. 1 we compare the analytic

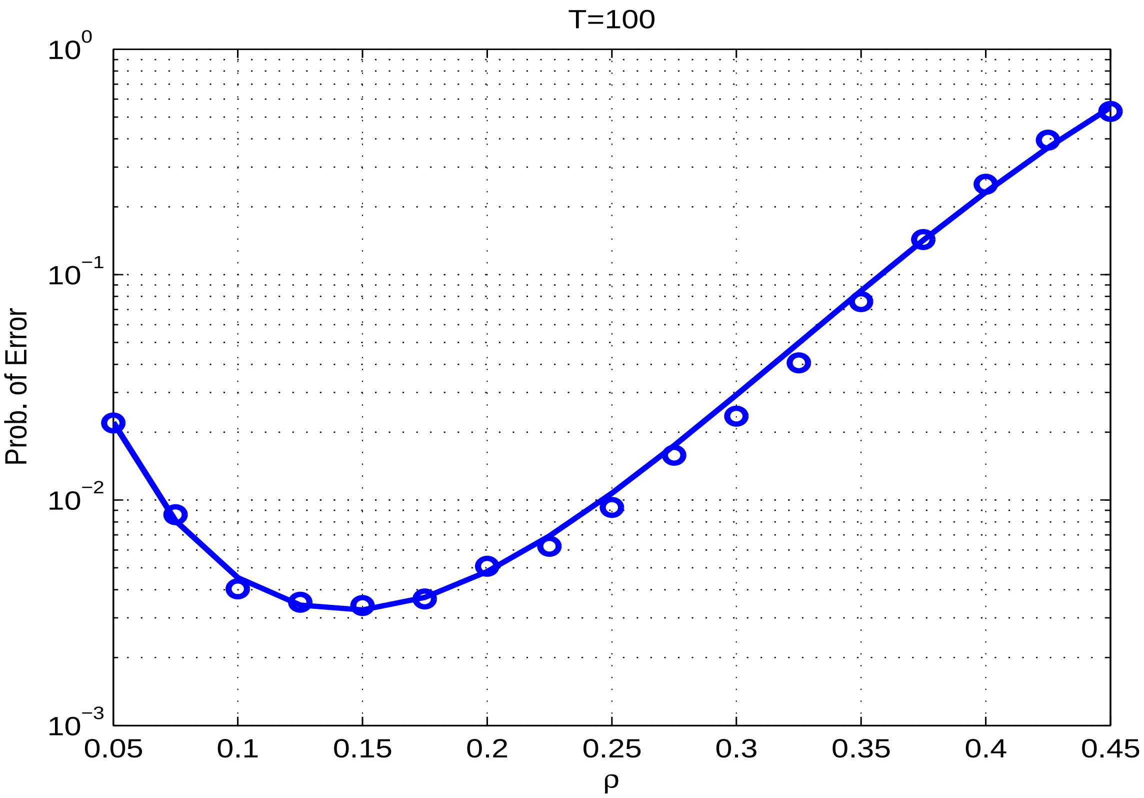

Fig. 1. Empirical probability of failure ('o') and its analytic approximation (solid) vs. the probability ρ for $P = 2$, $K = 2$ sources, $T = 100$. The empirical probabilities were obtained using $25,000$ independent trials

prediction (40) to the empirical probability of failure obtained in $25,000$ independent experiments (the sources and the mixing matrix were drawn independently in each trial) vs. ρ for $T = 100$. Failure of the separation is defined as the case in which $\hat{\boldsymbol{B}} \circ \boldsymbol{A}$ is not a permutation matrix. We used the AMERICA algorithm for separation (but for this ($K = 2$) case, similar results are obtained with MEXICO). The circles show the empirical probabilities, whereas the solid line shows the approximate analytic prediction (40). The good match is evident.

When K is larger than 2, an approximate error expression can be obtain by assuming that this type of error can occur independently for each of the $K(K-1)/2$ different couples. Under this approximate independence assumption, we get

$$\Pr\{\text{Failure}; K\} \approx 1 - (1 - \Pr\{\text{Failure}; K = 2\})^{K(K-1)}, \tag{41}$$

where $\Pr\{\text{Failure}; K = 2\}$ is given in (40) above. We assume here, for simplicity of the exposition, that all of the sources take the value 1 with similar probability ρ. Extension to the case of different probabilities can be readily obtained by using (39) for each (ordered) couple.

To illustrate, we compare this expression in Fig. 2 to the empirical probability of failure (obtained

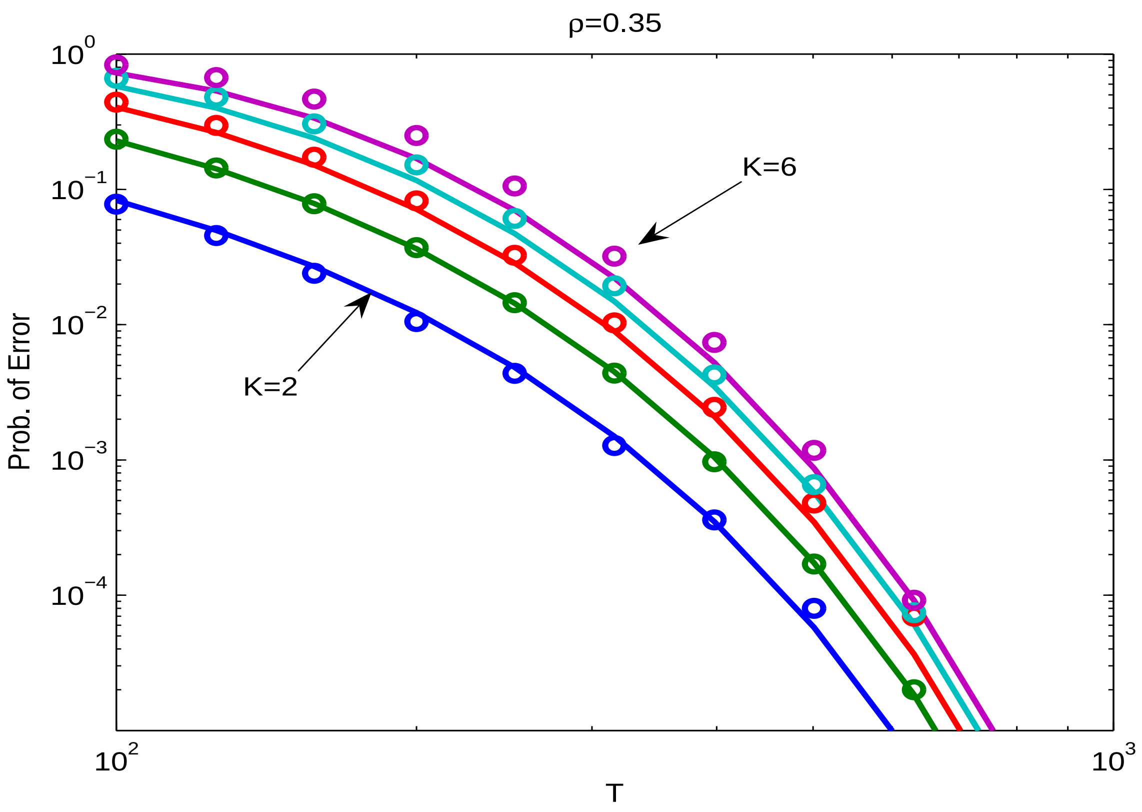

Fig. 2. Empirical probability of failure (’o’) and its analytic approximation (solid) vs. the observation length T for $P = 2$, $K = 2, 3, 4, 5$ and 6 sources with $\rho = 0.35$, using the AMERICA algorithm. The empirical probabilities were obtained using $100,000$ independent trials

in $100,000$ independent experiments) vs. T for $\rho = 0.35$ with $K = 2, 3, 4, 5, 6$. Again, failure of the separation is defined as the case in which $\hat{\boldsymbol{B}} \circ \boldsymbol{A}$ is not a permutation matrix (namely, any result which does not provide perfect separation of *all* of the K sources is considered a “failure”). A good match is evident for the smaller values of K, with some departure for the higher values - as could be expected from the approximation induced by the error-independence assumption.

Next, we compare the empirical, average running-times of the two separation algorithm under asymptotic conditions. The “asymptotic” conditions are emulated by substituting the estimated (empirical) probabilities tensor $\widehat{\boldsymbol{\mathcal{P}}}_{\boldsymbol{x}}$ with the true probabilities tensor $\boldsymbol{\mathcal{P}}_{\boldsymbol{x}}$ as the input to the algorithms. We simulated two cases: A “full” mixing matrix and a “sparse” mixing matrix. The “full” $K \times K$ (non-singular) mixing matrices were randomly drawn in each trial as a product of a lower triangular and an upper triangular matrix. The lower triangular matrix $\boldsymbol{L}$ was generated with random values independently and uniformly distributed in $\mathbb{GF}(P)$ on and below the main diagonal, substituting any 0-s along the main diagonal with 1-s; The upper diagonal matrix $\boldsymbol{U}$ was similarly generated by drawing all values above the main diagonal, and setting the main diagonal to all-1-s. Then $\boldsymbol{A} = \boldsymbol{U} \circ \boldsymbol{L}$. For

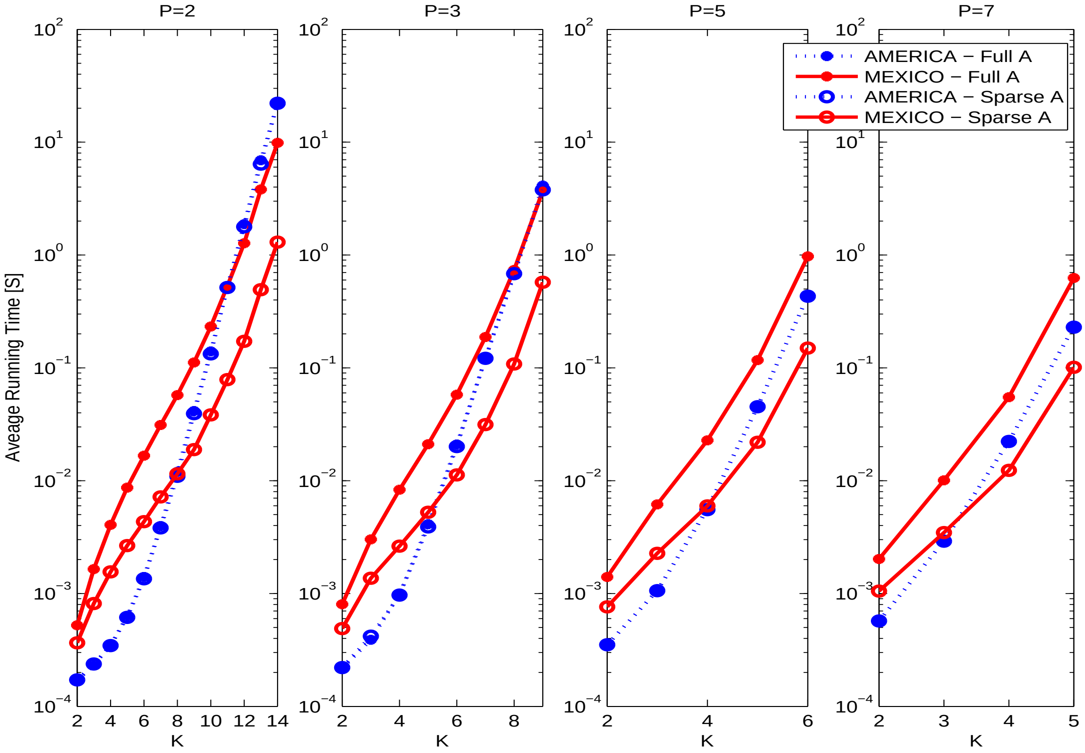

Fig. 3. Average running times (in [seconds]) for the AMERICA (dashed) and MEXICO (solid) algorithms, for full ('*') and sparse ('o') matrices. Note that the AMERICA plots for both the full and the sparse mixing case are nearly identical.

generating the "sparse" matrices, the off-diagonal values of $\boldsymbol{L}$ and $\boldsymbol{U}$ were "sparsified" by randomly (and independently) zeroing-out each element, with probability 0.9.

The elements of each of the sources' probabilities vectors $\boldsymbol{p}_{s_1}, \ldots \boldsymbol{p}_{s_K}$ were drawn uniformly in $(0,1)$ and then normalized by their sum. The average running times (using Matlab® code [11] for both algorithms on a PC Pentium® 4 running at 3.4GHz) for several combinations of P and K are shown in Fig. 3. Both algorithms were applied to the same data, and the running times were averaged over 4000 independent trials. As expected, the AMERICA algorithm is seen to be insensitive to the structure (full / sparse) of the mixing matrix; However, the MEXICO algorithm runs considerably faster when $\boldsymbol{A}$ is sparse. Therefore, in terms of running speed, MEXICO may be preferable when the mixing matrix is known to be sparse, especially for relatively high values of K.

Note, however, that this advantage is somewhat overcast by a degradation in the resulting separation performance. While perfect separation was obtained (thanks to the "asymptotic" conditions) in all of the timing experiments by the AMERICA algorithm, few cases of imperfect separation by MEXICO were encountered, especially in the highest values of K with the "full" mixtures.

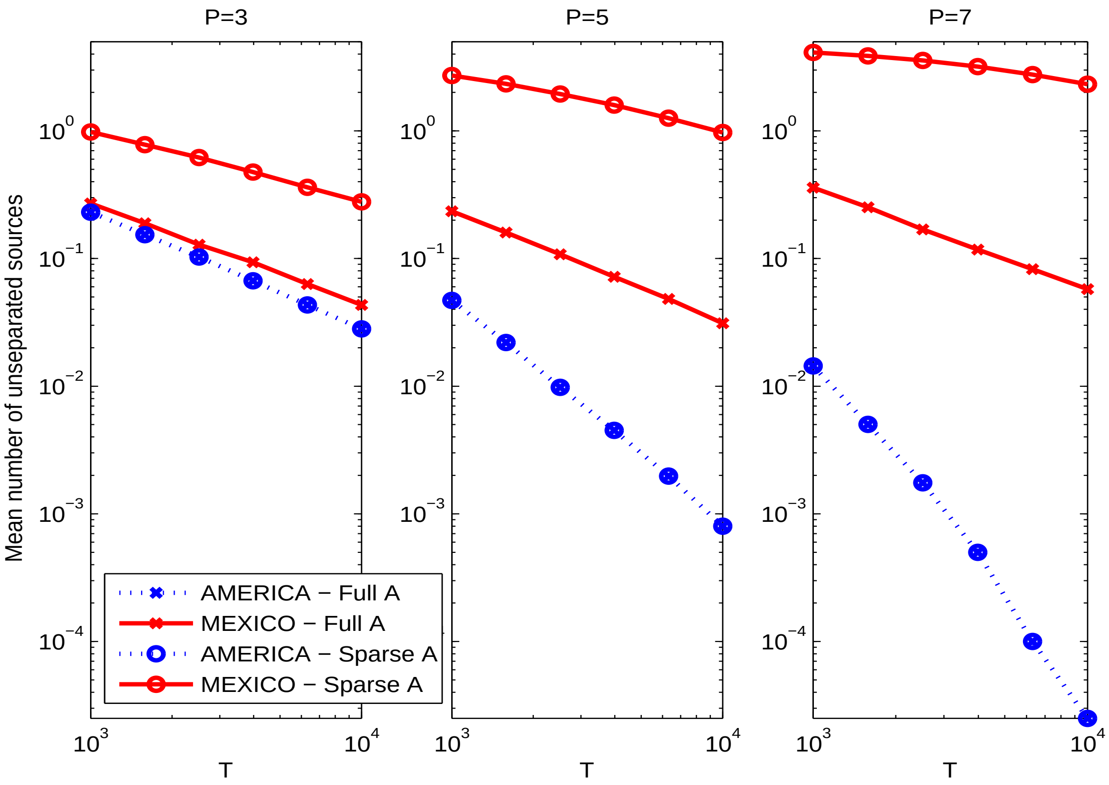

Fig. 4. Empirical mean number of unseparated sources (out of the $K = 5$ sources for AMERICA (dashed) and MEXICO (solid) algorithms, for full ('*') and sparse ('o') matrices for $P = 3, 5, 7$. Each point reflect the average of $40,000$ trials. Note that the AMERICA plots for both the full and the sparse mixing case are nearly identical.

To conclude this section, we provide (in Fig.4) some empirical results showing the performance for $P = 3, 5, 7$ with $K = 5$ sources, with random sources' probabilities vectors. The randomized elements of the probability vectors were independently drawn (for each source, at each trial) from a uniform distribution, and then normalized such that the sum of elements of each probability vector adds up to 1. The mixing matrix was randomized at each trial as described above, once for a "full $\boldsymbol{A}$" and once for a "sparse $\boldsymbol{A}$" version. In this experiment the performance is measured as the mean number of unseparated sources, which is defined (per trial) as the number of rows in the resulting "contamination matrix" $\hat{\boldsymbol{B}} \circ \boldsymbol{A}$ containing more than one non-zero element (since, by construction of $\hat{\boldsymbol{B}}$ in both MEXICO and AMERICA, $\hat{\boldsymbol{B}} \circ \boldsymbol{A}$ is always nonsingular, this is exactly the number of sources which remain unseparated by the algorithm). Each result on the plot reflects the average of $40,000$ trials.

Evidently, the AMERICA algorithm seems significantly more successful than the MEXICO algorithm, especially with the higher values of P (interestingly, the performance of AMERICA seems to improve with the increase in P, whereas the performance of MEXICO exhibits an opposite trend). The advantage of MEXICO is confined to cases of small P and large K, where its potentially reduced computational load does not come at the expense of a severe degradation in performance.

VII. Conclusion

We provided a study of general properties, identifiability conditions and separation algorithms for ICA over Galois fields of prime order P. We have shown that a linear mixture of independent sources is identifiable (up to permutation and, for $P > 2$, up to scale) if and only if none of the sources is uniform. We have shown that pairwise independence of an invertible linear mixture of the sources implies their full independence (namely, implies that the mixture is a scaled permutation) for $P = 2$ and for $P = 3$, but not necessarily for $P > 3$.

We proposed two different iterative separation algorithms: The first algorithm, given the acronym AMERICA, is based on sequential identification of the smallest-entropy linear combinations of the mixtures. The second, given the acronym MEXICO, is based on sequential reduction of the pairwise mutual information measures. We provided a rudimentary performance analysis for $P = 2$, which applies to both algorithms with $K = 2$, demonstrating a good fit of the empirical results to the theoretical prediction. For higher values of K (still with $P = 2$), we demonstrated a reasonable fir up to $K \approx 6$ for the AMERICA algorithm.

AMERICA is guaranteed to provide consistent separation (i.e., to recover all sources when the observation length T is infinite), and generally exhibits better performance (success rate) than MEXICO with finite data lengths. However, when the mixing-matrix is known to be sparse, MEXICO can have some advantage over AMERICA is in its relative computational efficiency, especially for larger values of K. Matlab® code for both algorithms is available online [11].

Extensions of our results to common variants of the classical ICA problem, such as ICA in the presence of additive noise, the under-determined case (more sources than mixtures), possible alternative sources of diversity (e.g., different temporal structures) of the sources, etc. - are all possible. For example, just like in classical ICA, temporal or spectral diversity would enable to relax the identifiability condition, so as to accommodate sources with uniform (marginal) distributions, which might be more commonly encountered. However, these extensions fall beyond the scope of our current work, whose main goal is to set the basis for migrating ICA from the real- (or complex-) valued algebraic fields to another.

Acknowledgement

The author would like to thank Jacob Goldberger for inspiring discussions on the use of FFT for the AMERICA algorithm.

Appendix A - A proof of Theorem 2

In this Appendix we provide a proof of Theorem 2 for both $\mathbb{GF}(2)$ and $\mathbb{GF}(3)$. Let $\boldsymbol{s}$ be a $K \times 1$ random vector whose elements are statistically-independent, non-degenerate and non-uniform random variables in either $\mathbb{GF}(2)$ or $\mathbb{GF}(3)$, and let $\boldsymbol{y} = \boldsymbol{D} \circ \boldsymbol{s}$ denote a $K \times 1$ vector of non-trivial linear combinations of the elements of $\boldsymbol{s}$ over the field, prescribed by the elements of the $K \times K$ matrix $\boldsymbol{D}$ (in either $\mathbb{GF}(2)$ or $\mathbb{GF}(3)$, resp.).

Assume that $\boldsymbol{D}$ is a general matrix, and consider any pair y_k and y_ℓ ($k \neq \ell$) in $\boldsymbol{y}$. y_k and y_ℓ are linear combinations of respective groups of the sources, indexed by the non-zero elements in $\boldsymbol{D}_{k,:}$ and $\boldsymbol{D}_{\ell,:}$, the k-th and ℓ-th rows (resp.) of $\boldsymbol{D}$.

Let us consider the case of $\mathbb{GF}(2)$ first.

A. *The* $\mathbb{GF}(2)$ *case*

The two groups composing y_k and y_ℓ define, in turn, three other subgroups (some of which may be empty):

1) Sub-group 1: Sources common to $\boldsymbol{D}_{k,:}$ and $\boldsymbol{D}_{\ell,:}$. Denote the sum of these sources as u;
2) Sub-group 2: Sources included in $\boldsymbol{D}_{k,:}$ but excluded from $\boldsymbol{D}_{\ell,:}$. Denote the sum of these sources as v_1;
3) Sub-group 3: Sources included in $\boldsymbol{D}_{\ell,:}$ but excluded from $\boldsymbol{D}_{k,:}$. Denote the sum of these sources as v_2.

For example, if (for $K = 6$) $\boldsymbol{D}_{k,:} = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$ and $\boldsymbol{D}_{\ell,:} = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 1 \end{bmatrix}$, then $u = s_2 \oplus s_5 \oplus s_6$, $v_1 = s_3 \oplus s_4$ and $v_2 = s_1$.

Note that by construction (and by independence of the elements of $\boldsymbol{s}$), the random variables u, v_1 and v_2 are statistically independent. Their respective probabilities vectors and characteristic vectors are denoted

$$\boldsymbol{p}_\nu = \begin{bmatrix} p_\nu(0) \\ p_\nu(1) \end{bmatrix} \ , \ \tilde{\boldsymbol{p}}_\nu = \begin{bmatrix} 1 \\ \theta_\nu \end{bmatrix} \ , \ \text{with} \ \ \theta_\nu = 1 - 2p_\nu(1) \ , \ \ \text{for} \quad \nu = u, v_1, v_2. \tag{42}$$

Obviously, $y_k = u \oplus v_1$ and $y_\ell = u \oplus v_2$, so their characteristic vectors are given by

$$\tilde{\boldsymbol{p}}_{y_k} = \tilde{\boldsymbol{p}}_u \odot \tilde{\boldsymbol{p}}_{v_1} = \begin{bmatrix} 1 \\ \theta_u \theta_{v_1} \end{bmatrix} \ , \ \tilde{\boldsymbol{p}}_{y_\ell} = \tilde{\boldsymbol{p}}_u \odot \tilde{\boldsymbol{p}}_{v_2} = \begin{bmatrix} 1 \\ \theta_u \theta_{v_2} \end{bmatrix}, \tag{43}$$

where $\odot$ denotes the Hadamard (element-wise) product.

Define the random vector $\boldsymbol{w} \triangleq [y_k \ y_\ell]^T$, which can be expressed as the sum of three independent random vectors:

$$\underbrace{\begin{bmatrix} y_k \\ y_\ell \end{bmatrix}}_{\boldsymbol{w}} = \underbrace{\begin{bmatrix} v_1 \\ 0 \end{bmatrix}}_{\triangleq \boldsymbol{v}_1} \oplus \underbrace{\begin{bmatrix} u \\ u \end{bmatrix}}_{\triangleq \boldsymbol{u}} \oplus \underbrace{\begin{bmatrix} 0 \\ v_2 \end{bmatrix}}_{\triangleq \boldsymbol{v}_2} \tag{44}$$

The probabilities matrices of the vectors $\boldsymbol{v}_1$, $\boldsymbol{v}_2$ and $\boldsymbol{u}$ are evidently given by

$$\boldsymbol{P}_{\boldsymbol{v}_1} = \begin{bmatrix} 1-p_{v_1}(1) & 0 \\ p_{v_1}(1) & 0 \end{bmatrix} \quad \boldsymbol{P}_{\boldsymbol{v}_2} = \begin{bmatrix} 1-p_{v_2}(1) & p_{v_2}(1) \\ 0 & 0 \end{bmatrix} \quad \boldsymbol{P}_{\boldsymbol{u}} = \begin{bmatrix} 1-p_{u}(1) & 0 \\ 0 & p_{u}(1) \end{bmatrix} \tag{45}$$

and therefore their characteristic matrices are given by

$$\widetilde{\boldsymbol{P}}_{\boldsymbol{v}_1} = \begin{bmatrix} 1 & 1 \\ \theta_{v_1} & \theta_{v_1} \end{bmatrix} \quad \widetilde{\boldsymbol{P}}_{\boldsymbol{v}_2} = \begin{bmatrix} 1 & \theta_{v_2} \\ 1 & \theta_{v_2} \end{bmatrix} \quad \widetilde{\boldsymbol{P}}_{\boldsymbol{u}} = \begin{bmatrix} 1 & \theta_{u} \\ \theta_{u} & 1 \end{bmatrix}, \tag{46}$$

where (see Section II) $\theta_\nu = E[W_2^\nu] = E[(-1)^\nu] = 1 - 2p_\nu(1)$, for $\nu = v_1, v_2, u$. Since $\boldsymbol{v}_1$, $\boldsymbol{v}_2$ and $\boldsymbol{u}$ are statistically independent, the characteristic matrix of $\boldsymbol{w}$ is given by

$$\widetilde{\boldsymbol{P}}_{\boldsymbol{w}} = \widetilde{\boldsymbol{P}}_{\boldsymbol{v}_1} \odot \widetilde{\boldsymbol{P}}_{\boldsymbol{u}} \odot \widetilde{\boldsymbol{P}}_{\boldsymbol{v}_1} = \begin{bmatrix} 1 & \theta_u\theta_{v_2} \\ \theta_{v_1}\theta_u & \theta_{v_1}\theta_{v_2} \end{bmatrix}. \tag{47}$$

On the other hand, if y_k and y_ℓ are statistically independent, the characteristic matrix of $\boldsymbol{w}$ is also given by

$$\widetilde{\boldsymbol{P}}_{\boldsymbol{w}} = \tilde{\boldsymbol{p}}_{y_k}\tilde{\boldsymbol{p}}_{y_\ell}^T = \begin{bmatrix} 1 & \theta_{v_2}\theta_u \\ \theta_u\theta_{v_2} & \theta_u^2\theta_{v_1}\theta_{v_2} \end{bmatrix}. \tag{48}$$

Equating the expressions on (47) and (48), we get (only the $(2,2)$ element can differ)

$$\theta_u^2\theta_{v_1}\theta_{v_2} = \theta_{v_1}\theta_{v_2}. \tag{49}$$

Since, due to Lemma 1, if neither of the sources is uniform, neither are v_1 and v_2, we have $\theta_{v_1}, \theta_{v_2} \neq 0$, and therefore θ_u must be either 1 or -1. Since neither of the sources is degenerate, this can only happen if $u = 0$ (deterministically), which can only happen if sub-group 1 is empty, namely, if the two rows $\boldsymbol{D}_{k,:}$ and $\boldsymbol{D}_{\ell,:}$ do not share common sources, or, in other words, if there is no column m in $\boldsymbol{D}$ such that both $\boldsymbol{D}_{k,m}$ and $\boldsymbol{D}_{\ell,m}$ are 1.

Applying this to all possible pairs of $k \neq \ell$ (for which y_k and y_ℓ are independent), and recalling that $\boldsymbol{D}$ cannot have any all-zeros row (no trivial combinations in $\boldsymbol{y}$), we immediately arrive at the conclusion that each row and each column of $\boldsymbol{D}$ must contain exactly one 1, meaning that $\boldsymbol{D}$ is a permutation matrix.

We now turn to the case of $\mathbb{GF}(3)$.

B. The $\mathbb{GF}(3)$ case

For simplicity of the exposition, we shall now assume that the values taken in $\mathbb{GF}(3)$ are $\{0, 1-1\}$ (rather than $\{0,1,2\}$). Just like in the $\mathbb{GF}(2)$ case, we partition the two groups composing y_k and y_ℓ into subgroups, but now the first ("common") subgroup is further partitioned into three sub-subgroups:

1) Sub-group 1: Sources common to $\boldsymbol{D}_{k,:}$ and $\boldsymbol{D}_{\ell,:}$. We partition this sub-group into four sub-subgroups according to the coefficients in the respective rows of $\boldsymbol{D}$ as follows:

- Sub-subgroup 1a: sources for which the respective coefficients in $\boldsymbol{D}_{k,:}$ and $\boldsymbol{D}_{\ell,:}$ are both 1; Denote the sum of these sources as u_{++};
- Sub-subgroup 1b: sources for which the respective coefficients in $\boldsymbol{D}_{k,:}$ and $\boldsymbol{D}_{\ell,:}$ are both -1; Denote the sum of these sources as u_{--};
- Sub-subgroup 1c: sources for which the respective coefficients in $\boldsymbol{D}_{k,:}$ and $\boldsymbol{D}_{\ell,:}$ are 1 and -1, resp.; Denote the sum of these sources as u_{+-};
- Sub-subgroup 1d: sources for which the respective coefficients in $\boldsymbol{D}_{k,:}$ and $\boldsymbol{D}_{\ell,:}$ are -1 and 1, resp.; Denote the sum of these sources as u_{-+};

2) Sub-group 2: Sources included in $\boldsymbol{D}_{k,:}$ but excluded from $\boldsymbol{D}_{\ell,:}$. Denote the respective linear combination of these sources as v_1;
3) Sub-group 3: Sources included in $\boldsymbol{D}_{\ell,:}$ but excluded from $\boldsymbol{D}_{k,:}$. Denote the respective linear combination of these sources as v_2.

For example, if (for $K=6$) $\boldsymbol{D}_{k,:} = \begin{bmatrix} 0 & 1 & -1 & 1 & 1 & 1 \end{bmatrix}$ and $\boldsymbol{D}_{\ell,:} = \begin{bmatrix} -1 & -1 & 0 & 0 & 1 & 1 \end{bmatrix}$, then $u_{++} = s_5 \oplus s_6$, $u_{--} = u_{-+} = 0$, $u_{+-} = s_2$, $v_1 = -s_3 \oplus s_4$ and $v_2 = -s_1$.

The random variables u_{++}, u_{--}, u_{+-}, u_{-+}, v_1 and v_2 are statistically independent. Their respective probabilities vectors and characteristic vectors are denoted

$$\boldsymbol{p}_\nu = \begin{bmatrix} p_\nu(0) \\ p_\nu(1) \\ p_\nu(2) \end{bmatrix} \;,\; \tilde{\boldsymbol{p}}_\nu = \begin{bmatrix} 1 \\ \xi_\nu \\ \xi_\nu^* \end{bmatrix} \;,\; \text{for} \quad \nu = u_{++}, u_{--}, u_{+-}, u_{-+}, v_1, v_2. \tag{50}$$

An expression for $\xi_\nu = E[W_3^\nu]$ in terms of $p_\nu(0)$, $p_\nu(1)$ and $p_\nu(2)$ can be found in (5) above. Note further, that $\xi_{-\nu} = \xi_\nu^*$, so that $\tilde{\boldsymbol{p}}_{-\nu} = \tilde{\boldsymbol{p}}_\nu^*$.

Evidently,

$$y_k = v_1 \oplus u_{++} \ominus u_{--} \oplus u_{+-} \ominus u_{-+}; \;,\; y_\ell = v_2 \oplus u_{++} \ominus u_{--} \ominus u_{+-} \oplus u_{--}, \tag{51}$$

so their characteristic vectors are given by

$$\begin{aligned} \tilde{\boldsymbol{p}}_{y_k} &= \tilde{\boldsymbol{p}}_{v_1} \odot \tilde{\boldsymbol{p}}_{u_{++}} \odot \tilde{\boldsymbol{p}}_{u_{--}}^* \odot \tilde{\boldsymbol{p}}_{u_{+-}} \odot \tilde{\boldsymbol{p}}_{u_{-+}}^* \\ \tilde{\boldsymbol{p}}_{y_\ell} &= \tilde{\boldsymbol{p}}_{v_2} \odot \tilde{\boldsymbol{p}}_{u_{++}} \odot \tilde{\boldsymbol{p}}_{u_{--}}^* \odot \tilde{\boldsymbol{p}}_{u_{+-}}^* \odot \tilde{\boldsymbol{p}}_{u_{-+}} \end{aligned} \tag{52}$$

The random vector $\boldsymbol{w} \triangleq [y_k \; y_\ell]^T$ can now be expressed as the sum of five independent random vectors:

$$\underbrace{\begin{bmatrix} y_k \\ y_\ell \end{bmatrix}}_{\boldsymbol{w}} = \underbrace{\begin{bmatrix} v_1 \\ 0 \end{bmatrix}}_{\triangleq \boldsymbol{v}_1} \oplus \underbrace{\begin{bmatrix} u_{++} \\ u_{++} \end{bmatrix}}_{\triangleq \boldsymbol{u}_{++}} \oplus \underbrace{\begin{bmatrix} -u_{--} \\ -u_{--} \end{bmatrix}}_{\triangleq \boldsymbol{u}_{--}} \oplus \underbrace{\begin{bmatrix} u_{+-} \\ -u_{+-} \end{bmatrix}}_{\triangleq \boldsymbol{u}_{+-}} \oplus \underbrace{\begin{bmatrix} -u_{-+} \\ u_{-+} \end{bmatrix}}_{\triangleq \boldsymbol{u}_{-+}} \oplus \underbrace{\begin{bmatrix} 0 \\ v_2 \end{bmatrix}}_{\triangleq \boldsymbol{v}_2} \tag{53}$$

The probabilities matrices of the vectors $\boldsymbol{v}_1$, $\boldsymbol{v}_2$, $\boldsymbol{u}_{++}$, $\boldsymbol{u}_{--}$, $\boldsymbol{u}_{+-}$ and $\boldsymbol{u}_{-+}$, and their respective characteristic matrices are given by

$$\boldsymbol{P}_{\boldsymbol{v}_1} = \begin{bmatrix} p_{v_1}(0) & 0 & 0 \\ p_{v_1}(1) & 0 & 0 \\ p_{v_1}(2) & 0 & 0 \end{bmatrix} \Rightarrow \widetilde{\boldsymbol{P}}_{\boldsymbol{v}_1} = \begin{bmatrix} 1 & 1 & 1 \\ \xi_{v_1} & \xi_{v_1} & \xi_{v_1} \\ \xi^*_{v_1} & \xi^*_{v_1} & \xi^*_{v_1} \end{bmatrix}; \tag{54a}$$

$$\boldsymbol{P}_{\boldsymbol{v}_2} = \begin{bmatrix} p_{v_2}(0) & p_{v_2}(1) & p_{v_2}(2) \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \Rightarrow \widetilde{\boldsymbol{P}}_{\boldsymbol{v}_1} = \begin{bmatrix} 1 & \xi_{v_2} & \xi^*_{v_2} \\ 1 & \xi_{v_2} & \xi^*_{v_2} \\ 1 & \xi_{v_2} & \xi^*_{v_2} \end{bmatrix}; \tag{54b}$$

$$\boldsymbol{P}_{\boldsymbol{u}_{++}} = \begin{bmatrix} p_{u_{++}}(0) & 0 & 0 \\ 0 & p_{u_{++}}(1) & 0 \\ 0 & 0 & p_{u_{++}}(2) \end{bmatrix} \Rightarrow \widetilde{\boldsymbol{P}}_{\boldsymbol{u}_{++}} = \begin{bmatrix} 1 & \xi_{u_{++}} & \xi^*_{u_{++}} \\ \xi_{u_{++}} & \xi^*_{u_{++}} & 1 \\ \xi^*_{u_{++}} & 1 & \xi_{u_{++}} \end{bmatrix}; \tag{55a}$$

$$\boldsymbol{P}_{\boldsymbol{u}_{--}} = \begin{bmatrix} p_{u_{--}}(0) & 0 & 0 \\ 0 & p_{u_{--}}(2) & 0 \\ 0 & 0 & p_{u_{--}}(1) \end{bmatrix} \Rightarrow \widetilde{\boldsymbol{P}}_{\boldsymbol{u}_{--}} = \begin{bmatrix} 1 & \xi^*_{u_{--}} & \xi_{u_{--}} \\ \xi^*_{u_{--}} & \xi_{u_{--}} & 1 \\ \xi_{u_{--}} & 1 & \xi^*_{u_{--}} \end{bmatrix}; \tag{55b}$$

$$\boldsymbol{P}_{\boldsymbol{u}_{+-}} = \begin{bmatrix} p_{u_{+-}}(0) & 0 & 0 \\ 0 & 0 & p_{u_{+-}}(1) \\ 0 & p_{u_{+-}}(2) & 0 \end{bmatrix} \Rightarrow \widetilde{\boldsymbol{P}}_{\boldsymbol{u}_{+-}} = \begin{bmatrix} 1 & \xi^*_{u_{+-}} & \xi_{u_{+-}} \\ \xi_{u_{+-}} & 1 & \xi^*_{u_{+-}} \\ \xi^*_{u_{+-}} & \xi_{u_{+-}} & 1 \end{bmatrix}; \tag{55c}$$

$$\boldsymbol{P}_{\boldsymbol{u}_{-+}} = \begin{bmatrix} p_{u_{-+}}(0) & 0 & 0 \\ 0 & 0 & p_{u_{-+}}(2) \\ 0 & p_{u_{-+}}(1) & 0 \end{bmatrix} \Rightarrow \widetilde{\boldsymbol{P}}_{\boldsymbol{u}_{+-}} = \begin{bmatrix} 1 & \xi_{u_{-+}} & \xi^*_{u_{-+}} \\ \xi^*_{u_{-+}} & 1 & \xi_{u_{-+}} \\ \xi_{u_{-+}} & \xi^*_{u_{-+}} & 1 \end{bmatrix}; \tag{55d}$$

Thus, the characteristic matrix of $\boldsymbol{w}$ is given by the Hadamard product of these matrices,

$$\widetilde{\boldsymbol{P}}_{\boldsymbol{w}} = \widetilde{\boldsymbol{P}}_{v_1} \odot \widetilde{\boldsymbol{P}}_{v_2} \odot \widetilde{\boldsymbol{P}}_{u_{++}} \odot \widetilde{\boldsymbol{P}}_{u_{--}} \odot \widetilde{\boldsymbol{P}}_{u_{+-}} \odot \widetilde{\boldsymbol{P}}_{u_{-+}}. \tag{56}$$

Now, if y_k and y_ℓ are statistically independent, then $\widetilde{\boldsymbol{P}}_{\boldsymbol{w}}$ is also given by the outer product of their characteristic vectors, which, using (52), is given by

$$\widetilde{\boldsymbol{P}}_{\boldsymbol{w}} = \tilde{\boldsymbol{p}}_{y_k}\tilde{\boldsymbol{p}}_{y_\ell}^T = (\tilde{\boldsymbol{p}}_{v_1}\tilde{\boldsymbol{p}}_{v_2}^T) \odot (\tilde{\boldsymbol{p}}_{u_{++}}\tilde{\boldsymbol{p}}_{u_{++}}^T) \odot (\tilde{\boldsymbol{p}}^*_{u_{--}}\tilde{\boldsymbol{p}}^H_{u_{--}}) \odot (\tilde{\boldsymbol{p}}_{u_{+-}}\tilde{\boldsymbol{p}}^H_{u_{+-}}) \odot (\tilde{\boldsymbol{p}}^*_{u_{-+}}\tilde{\boldsymbol{p}}^T_{u_{-+}}), \tag{57}$$

where $(\cdot)^H$ denotes the conjugate transpose. Noting that $\tilde{\boldsymbol{p}}_{v_1}\tilde{\boldsymbol{p}}_{v_2}^T = \widetilde{\boldsymbol{P}}_{v_1} \odot \widetilde{\boldsymbol{P}}_{v_2}$, and recalling that, since v_1 and v_2 cannot be uniform, ξ_{v_1} and ξ_{v_2} must be non-zero, we conclude that the independence of y_k and y_ℓ implies that

$$(\tilde{\boldsymbol{p}}_{u_{++}}\tilde{\boldsymbol{p}}_{u_{++}}^T) \odot (\tilde{\boldsymbol{p}}^*_{u_{--}}\tilde{\boldsymbol{p}}^H_{u_{--}}) \odot (\tilde{\boldsymbol{p}}_{u_{+-}}\tilde{\boldsymbol{p}}^H_{u_{+-}}) \odot (\tilde{\boldsymbol{p}}^*_{u_{-+}}\tilde{\boldsymbol{p}}^T_{u_{-+}}) = \widetilde{\boldsymbol{P}}_{u_{++}} \odot \widetilde{\boldsymbol{P}}_{u_{--}} \odot \widetilde{\boldsymbol{P}}_{u_{+-}} \odot \widetilde{\boldsymbol{P}}_{u_{-+}}. \tag{58}$$

It is easy to observe, that the first row and first column of each of the matrices on the left-hand side (LHS) are indeed always identical to those of the respective matrices on the right-hand side (RHS), regardless of the values of the ξ parameters. In addition, in each of the matrices the $(2,2)$ element[4] is the conjugate of the $(3,3)$ element, and the $(2,3)$ element is the conjugate of the $(3,2)$ element. Therefore, the independence of y_k and y_ℓ merely implies the equality of the products of the $(2,2)$ elements on the LHS and on the RHS, and of the products of the $(2,3)$ elements on the LHS and on the RHS.

The equality of the product of the $(2,2)$ elements implies

$$\xi^*_{u_{++}} \cdot \xi_{u_{--}} \cdot 1 \cdot 1 = (\xi_{u_{++}})^2 \cdot (\xi^*_{u_{--}})^2 \cdot |\xi_{u_{+-}}|^2 \cdot |\xi_{u_{-+}}|^2, \tag{59a}$$

and the equality of the product of the $(2,3)$ elements implies

$$1 \cdot 1 \cdot \xi^*_{u_{+-}} \cdot \xi_{u_{-+}} = |\xi_{u_{++}}|^2 \cdot |\xi_{u_{--}}|^2 \cdot (\xi_{u_{+-}})^2 \cdot (\xi^*_{u_{-+}})^2. \tag{59b}$$

Taking the absolute values of both, and recalling that since neither of the random variables u_{++}, u_{--}, u_{+-} and u_{-+} can be uniform, neither of the ξ parameters can be zero, we have

$$\begin{aligned}
|\xi_{u_{++}}| \cdot |\xi_{u_{--}}| = |\xi_{u_{++}}|^2 \cdot |\xi_{u_{--}}|^2 \cdot |\xi_{u_{+-}}|^2 \cdot |\xi_{u_{-+}}|^2 \;\Rightarrow\; |\xi_{u_{++}}| \cdot |\xi_{u_{--}}| \cdot |\xi_{u_{+-}}|^2 \cdot |\xi_{u_{-+}}|^2 = 1 \\
|\xi_{u_{+-}}| \cdot |\xi_{u_{-+}}| = |\xi_{u_{++}}|^2 \cdot |\xi_{u_{--}}|^2 \cdot |\xi_{u_{+-}}|^2 \cdot |\xi_{u_{-+}}|^2 \;\Rightarrow\; |\xi_{u_{++}}|^2 \cdot |\xi_{u_{--}}|^2 \cdot |\xi_{u_{+-}}| \cdot |\xi_{u_{-+}}| = 1.
\end{aligned} \tag{60}$$

Since for any random variable ν in $\mathbb{GF}(3)$, $|\xi_\nu| \leq 1$ with equality iff ν is degenerate, we conclude from (60) that if y_k and y_ℓ are independent, then u_{++}, u_{--}, u_{+-} and u_{-+} must all be degenerate. Since none of the independent sources is degenerate, this implies, in turn, that all four are identically zero, and that there are no non-zero elements common to $\boldsymbol{D}_{k,:}$ and $\boldsymbol{D}_{\ell,:}$.

Like in the $\mathbb{GF}(2)$ case, by repeated application of this result to all row-couples in $\boldsymbol{D}$, we conclude that pairwise independence of the elements of $\boldsymbol{y}$ implies that $\boldsymbol{D}$ is (up to signs) permutation matrix, namely that the elements of $\boldsymbol{y}$ are fully mutually independent.

REFERENCES

[1] A. Yeredor, "ICA in boolean XOR mixtures," *Proc., ICA 2007, Lecture Notes in Computer Science (LNCS 4666)*, vol. 45, no. 1, pp. 17–24, 2007.

[2] P. Comon, "Independent component analysis - a new concept?," *Signal Processing*, vol. 36, pp. 287–314, 1994.

[3] Jean-François Cardoso, "Blind signal separation: statistical principles," *Proceedings of the IEEE*, vol. 86, no. 10, pp. 2009–2025, 1998.

[4] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley & Sons, inc., 2001.

[5] J. Eriksson and V. Koivunen, "Complex random vectors and ICA models: identifiability, uniqueness, and separability," *IEEE Trans. Information Theory*, vol. 52, no. 3, pp. 1017–29, 2006.

[6] C.R. Rao, *Linear Statistical lnjerence and its Applications*, Wiley, New-York, 1973.

[7] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines, "A blind source separation technique using second-order statistics," *IEEE Trans. Signal Processing*, vol. 45, no. 2, pp. 434–44, 1997.

[4] In this context (only) we enumerate these matrices' rows and columns as $1,2,3$ rather than $0,1,2$.

[8] D.-T. Pham and J.-F. Cardoso, "Blind separation of instantaneous mixtures of nonstationary sources," *IEEE Trans. Signal Processing*, vol. 49, no. 9, pp. 1837–48, 2001.

[9] N. Delfosse and P. Loubaton, "Adaptive blind separation of independent sources: a deflation approach," *Signal Processing*, vol. 45, no. 1, pp. 59–83, 1995.

[10] J.-F. Cardoso and B. Laheld, "Equivariant adaptive source separation," *IEEE Trans. on Signal Processing*, vol. 44, no. 12, pp. 3017–3030, 1996.

[11] A. Yeredor, *Matlab® code for ICA over $\mathbb{GF}(P)$, AMERICA and MEXICO*,
Online: `http://www.eng.tau.ac.il/∼arie/ICA4GFP.rar`.